Ari Hartikainen

Statistical analysis of geological space

School of Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 26.5.2014

Thesis supervisor:

Prof. Jussi Leveinen

Thesis advisors:

Lic.Sc. (Phil.) Ursula Sievänen

D.Sc. (Phil.) Ari Luukkonen

Aalto University
School of Engineering

AALTO UNIVERSITY
SCHOOL OF ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Ari Hartikainen

Title: Statistical analysis of geological space

Date: 26.5.2014 Language: English Number of pages: 7+63

Department of Civil and Environmental Engineering

Professorship: Engineering Geology and Applied Geophysics Code: Yhd-33

Supervisor: Prof. Jussi Leveinen

Advisors: Lic.Sc. (Phil.) Ursula Sievänen, D.Sc. (Phil.) Ari Luukkonen

Statistical analyses of geological properties needs specific statistical tools. The tools used in this thesis includes descriptive statistics, hypothesis testing and visual presentations in forms of histograms, kernel density plots and quantile-quantile plots. Data are discretised by inverse distance procedure and they are processed with principal factor analysis and multiple regression analysis with stepwise backward model selection.

The variables chosen to be examined against transmissivity values are fracture frequency, single point resistance, fracture types and fracture filling including carbonates, sulphides and clays. The data are processed with the possible transformations including Box-Cox and square root procedures if visual inspection and descriptive statistics indicates distribution other than normal.

The principal factor analysis ends up with three principal factors where the first principal factor can be considered as the fragmentation of the bedrock. The second principal factor group describes electrical resistance relationship with lithology, sulphide and resistance values. Diatexitic gneiss and mica gneiss are found to correlate positively with higher resistance. The third factor appears as a subgroup for the first factor and indicates complex relationship between fracture surface filling with carbonate and the fracture types such as single fractures, hair dykes and single shears.

The multiple regression analysis gives insight into the relationship between the chosen variables and the transmissivity. It is found that the first factor and final regression model have similar properties. The lithological types including diatexitic and mica gneiss are found to have positively relationship with transmissivity. The carbonate variable does not survive in the regression and the reason for this is probably from the complex relationship with specific fracture types.

Keywords: Statistical analysis, Geology,
Transmissivity

AALTO-YLIOPISTO
INSINÖÖRITIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Ari Hartikainen

Työn nimi: Geologisen tilavuuden tilastollinen analyysi

Päivämäärä: 26.5.2014 Kieli: Englanti Sivumäärä: 7+63

Yhdyskunta- ja ympäristötekniikan laitos

Professuuri: Teknillinen geologia ja sovellettu geofysiikka Koodi: Yhd-33

Valvoja: Prof. Jussi Leveinen

Ohjaajat: FL Ursula Sievänen, FT Ari Luukkonen

Geologisen aineiston tilastollinen analyysi tarvitsee oikeanlaiset tilastolliset työkalut. Tässä diplomityössä käytetään tunnuslukuja ja visuaalisia keinoja kuvaamaan muuttujien jakaumaa. Tämän lisäksi hypoteesitestejä käytetään vahvistamaan pääteltyjä ominaisuuksia. Käytetty aineisto on sovitettu transmissiviteetin suhteen. Sovitetulle datalle tehdään pääkomponenttianalyysi sekä monta muuttujaa sisältävä lineaarinen regressiomalli takaisin askellus mallin avulla.

Muuttujat joiden suhdetta transmissiviteettiin halutaan tutkia on rakotiheys, pisteresistanssi, rakotyyppi sekä rakotäytemineralogia, joista on mukana karbonaatti, sulfidi sekä savi. Kerätylle ja sovitetulle datalle tehdään tarvittaessa Box-Cox tai neliöjuuri muunnos riippuen kuuluko data normaaliin jakaumaan vai mahdollisesti johonkin muuhun jakaumaan. Jakauman estimointi määritetään tunnuslukujen sekä visuaalisen aineiston avulla.

Pääkomponenttianalyysissa päädytään kolmeen pääkomponenttiin, joista kolmas pääkomponentti on ensimmäinen pääkomponenttiryhmän alaryhmä. Ensimmäisestä pääkomponenttia voidaan kuvailla kallion rikkonaisuudeksi. Toista pääkomponenttiryhmää voidaan kuvata sähkönjohtavuuden avulla. Tämän perusteella diateksiittinen gneissi ja kiillegneissi korreloivat positiivisesti korkean resistanssin kanssa. Kolmas pääkomponenttiryhmä kuvaa karbonaatin ja rakotyyppien monimutkaista suhdetta.

Lopullinen regressiomalli antaa selvemmän kuvan transmissiviteetin ja muiden muuttujien suhteesta toisiinsa. Regressiomalli antaa samankaltaisia tuloksia kuin pääkomponenttianalyysi, mutta selventää paremmin litologian ja transmissiviteetin suhdetta. Karbonaatti ei ole muuttujana lopullisessa mallissa, joka todennäköisesti johtuu rakotyypin ja karbonaatin monimutkaisesta suhteesta.

Avainsanat: Tilastollinen analyysi, Geologia, Transmissiviteetti

Preface

This thesis is done for the Radiation and Nuclear Safety Authority, Finland in winter 2013–2014. The conclusions and viewpoints presented in the thesis are those of author and do not necessarily coincide with those of Posiva.

I would like to thank my thesis supervisor professor Jussi Leveinen and my thesis advisors Ursula Sievänen and Ari Luukkonen for help and support.

Otaniemi, 26.5.2014

Ari Hartikainen

Contents

Symbols and abbreviations

Symbols

AIC	Akaike information criterian
IQR	interquartile range
$kurt$	kurtosis
max	sample maximum
min	sample minimum
Me	median
n	sample size
Q_j	jth quartile
s	corrected sample standard deviation
s^2	sample variance
$skew$	skewness
sup	suprematum of sample
$\overline{x}$	sample mean
β_j	jth regression coefficient
μ	population mean
σ^2	population variance

Abbreviations

CRUSH	crushed rock
DGN	diatexitic gneiss
Ft1	fracture surface type 1
Ft2	fracture surface type 2
Frac. freq	fracture frequency
KFP	K-feldspar porphyry
Lith.	lithology
MDB	diabase
MFGN	mafic gneiss
MGN	mica gneiss
ONKALO	underground rock characterisation facility at Olkiluoto
PGR	pegmatitic granite
Q-Q plot	quantile-quantile plot
QGN	quartz gneiss
SGN	stromatic gneiss
TGG	tonalitic-granodioritic-granitic gneiss
VGN	veined gneiss
$X_{\sqrt{}}$	square root transformation
$X_{log_{10}}$	logarithmic transformation in base 10

1 Introduction

Finland has four nuclear power plants in 2014. Two in west coast of Finland in municipality of Eurajoki and two in the south coast of Finland in city of Loviisa. One nuclear power plant is under construction at Olkiluoto. According to Ministry of Trade and Industry, Finland Nuclear Energy Act 990/1987 Section 6a:

"*Nuclear waste generated in connection with or as a result of use of nuclear energy in Finland shall be handled, stored and permanently disposed of in Finland.*"

Due to the Finnish legislation nuclear power plant owners Fortum Power and Heat and Teollisuuden Voima Oyj have surveyed different final disposal methods and decided to review more closely the deep geological repository option. In 1995 Fortum and Teollisuuden Voima established Posiva Oy to manage their used nuclear waste disposal. Finnish Parliament made decision-in-principle concerning final disposal of the nuclear waste, in December 2000. Based on investigations of Posiva, the Olkiluoto Island was selected to the site for further investigations in 2001. [Andersson et al., 2012]

Olkiluoto is an island located in municipality of Eurajoki, Finland at the coast of the Gulf of Bothnia. Posiva has been building an underground research facility ONKALO at Olkiluoto Island since 2004. The aim of the research facility is to study the local bedrock, to make a safety assessment for the disposal facility and test disposal techniques in real underground environment. Later the research facility is integrated to the final disposal facility. In deep geological repository concept the spent nuclear wastes is disposed to bedrock at depth of -400 metres and below. The current plan, based on the decision-in-principle, for disposal technique is to use KBS-3 method. In KBS-3 method spent fuel is sealed in water- and gas-tight copper canisters and placed in individual vertical disposal holes, KBS-3V, bored into the deposition tunnel. Alternative method places the disposal holes horizontally, KBS-3H. The copper canisters are installed within compressed clay elements that will after closure of repository swell as a release barrier. Finally deposition tunnel and central tunnels are backfilled with low-permeable material. The target for underground repository is to isolate disposed fuel from human environment safely the next 100 000 years in which time the activity of the nuclear fuel is drop to the natural levels [Hellä et al., 2013].

In the end of 2012 Posiva submitted a construction license application for the actual geological repository on the same site at Olkiluoto. The target for Posiva is to start the final disposal approximately on 2020. [Hellä et al., 2013]

Based on the Finnish Nuclear Energy Act 990/1987 Section 55, the Radiation and Nuclear Safety Authority, Finland are the supervising authority for the nuclear waste activities of Posiva and it inspects the license application. The supervising duties concern the safety of the spent nuclear fuels handling, short-term storage and final disposal.

The aim of this thesis is to examine potential statistical relationships with geological, hydrogeological and geophysical data within chosen space.

Data for analyses are constrained to the drillhole and –core studies made from the ground of Olkiluoto. The data are gathered from the POTTI database (POTTI

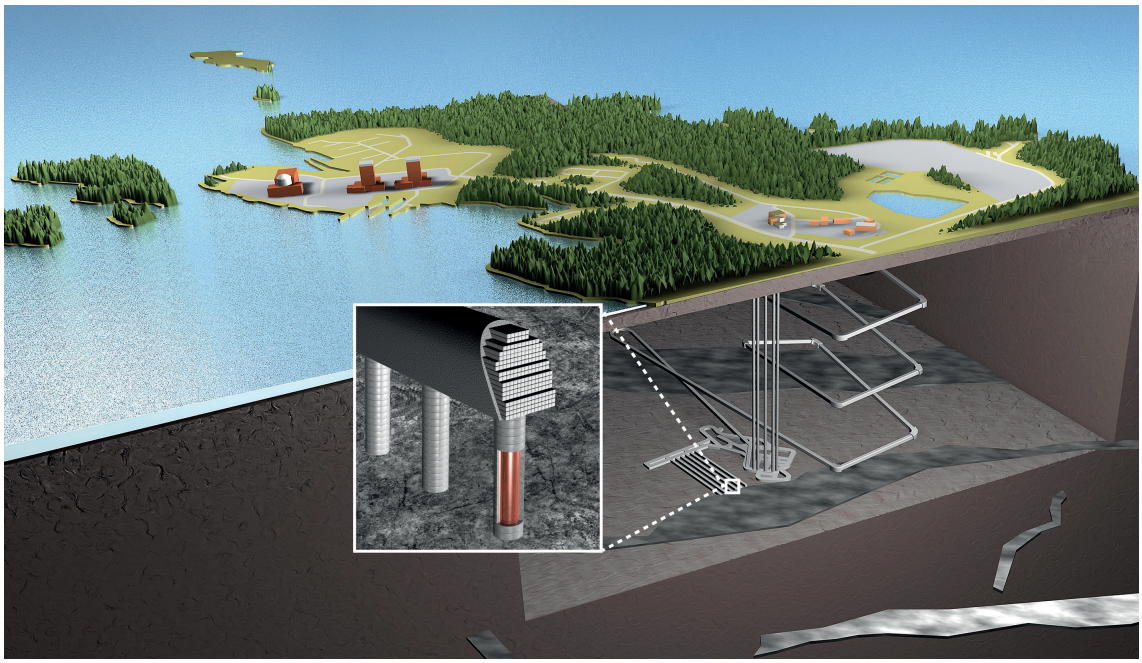

Figure 1.1: ONKALO Underground research facility and planned disposal repository with KBS-3V disposal holes. ©Posiva Oy.

is a OracleÂ® database with a browser-based interface, and is created for the administration of methods, studies and data of different disciplines in Posiva's site investigations) with 27.7.2013 as the date of retrieval. The data are composed of the geological, hydrogeological and geophysical studies gathered by Posiva. The main interest of analyses concerns relationships associated with fracturing and hydrogeological variables. Analyses start with inspection of basic statistical characteristics of the different data sets, which is described as descriptive statistics. The aim is to have a general knowledge and properties of the datasets and their distributions. After the descriptive statistics the data are resampled for the regression and principal component analyses. The data are resampled based on the dependent variable. The resampling for interval and discrete scales are made with the inverse distance weighting procedure and for categorical variables distance method is used. The categorical variables are studied and grouped to find out general relationship with each other. The resampled interval and discrete data and resampled and grouped categorical data are still refined for the regression and principal component analyses by trimming the non-complete groups. The regression and principal component analyses are performed with the refined datasets. The results are studied to find out possible trends.

This thesis is divided to four main parts. The first part of the thesis describes geological conditions, current state of the research made by the Posiva and defines available data. The second part manages statistical theory and procedures for analysing geological data. The third part of the thesis describes the data analyses procedures concerning chosen space and results. The fourth part offers discussion

and conclusion based on performed analyses and its results.

2 Current research and available data

2.1 Olkiluoto site

Geologically Olkiluoto site is a part of Satakunta suite [Aaltonen et al., 2010] and consist of two major lithological classes that are supracrustal high metamorphic rocks and igneous rocks. The first class includes migmatitic gneisses, tonalitic–granodioritic–granitic gneisses, mica gneisses, quartz gneisses and mafic gneisses. The second class consists of pegmatitic granites and diabase dykes. Posiva divides migmatitic gneisses to three subgroup that are veined gneisses, stromatic gneisses and diatexitic gneisses. Close to the Olkiluoto area there exists also a rapakivi batholith and to south of the rapakivi batholith a sandstone deposits. [Aaltonen et al., 2010].

2.2 Geological model 2.0

The latest published geological model of the Olkiluoto site is Geological model 2.0 [Aaltonen et al., 2010] with update in Andersson et al. [2012] Geological model 2.1, which is based on geological and geophysical information collected at Olkiluoto. It consists of four geological models that are the lithological site-scale model, the ductile deformation, the alteration and the brittle deformation model. The geological model considers connection between different models including the brittle deformation model and hydrogeology. The brittle deformation model considers geological zones that are product of truly brittle processes. These include joints, low-temperature mineral filled veins and fissures, faults and fault zones and diabase dykes. The ductile deformation model focuses on describing different non-brittle deformation phases that have occurred at the Olkiluoto site. The number of recognisable deformation phases are five. The fifth submodel belonging to the geological description is discrete fracture network model that is described in the Olkiluoto site description made by Andersson et al. [2012]. The Geological model 2.0 is described in the following subsections.

2.2.1 Lithological model

The lithological model is composed of 2D– and 3D–models. The 2D–model is a bedrock map with lithological rock types interpreted based on the outcrop observations and geophysical measurements.

The 3D lithological site–scale model consist of multiple modelled lithological units that are constructed based on the coarsened lithological data. In the coarsening procedure lithological observations from the drillcores longer than 10 metres in length are recognised as a separate lithological units. The pegmatitic granite observations less than this that are separated by short sections of gneiss are modelled as one larger unit of pegmatitic granites with the assumption that pegmatites are transecting dykes inside the gneiss units. The diatexitic gneiss is modelled as two large bodies in the south part of the modelling area and all the rest observations are discarded to veined gneiss due to lack of clear connection with each other. All the

diabase observations are modelled in spite of their size. Quartz gneiss observations have not been modelled due to their small size and all of the K-feldspar sections have been disregarded due to missing orientation data. Everything else that is not modelled as their own lithological units are marked as background comprising veined gneiss. The general direction of the lithological units are modelled as dipping to south to south east direction with moderate dip angle.

Diatexitic gneisses are modelled with two units that are located in the south part of the Olkiluoto area. Most diatexitic gneisses have migmatitic structure. The other occurrances of the diatexitic gneisses are merged to veined gneiss as explained in the previous paragraph.

Modelled mica gneisses consist of 41 individual units and are mostly scattered to the north from the diatexitic gneiss units. Mafic gneisses consist of four individual units scattered in the volume north of diatexitic gneiss units. Both mica gneisses and mafic gneisses are modelled as small lenses and have likely slightly higher fracture frequency.

Tonalitic-granodioritic-granitic gneisses consist of 52 individual units that are mostly homogenous and weakly fractured. North part of the modelling volume includes large units of tonalitic-granodioritic-granitic gneisses. Some of the modelled units are inside of the diatexitic gneiss units. 3D–modelled tonalitic-granodioritic-granitic gneisses are presented in Figure 2.1 as yellow solids.

Pegmatitic granites include 96 different units. Modelled pegmatitic granites are assumed to be heterogeneous based on the observations in outcrops and investigation trenches and further heterogeneity is included in coarsening procedure where nearby pegmatitic granite observations are merged together with gneiss inclusions. In the Geological model 2.0, pegmatite units are not assumed to reflect real shape of the dykes but are representations of volumes that have more pegmatitic granites than usual. In the model report it is emphasised that the modelled pegmatitic granite units have been formed at least in two deformation phases that adds uncertainty to the model. The 3D–modelled pegmatitic granites are presented in Figure 2.1 as red solids.

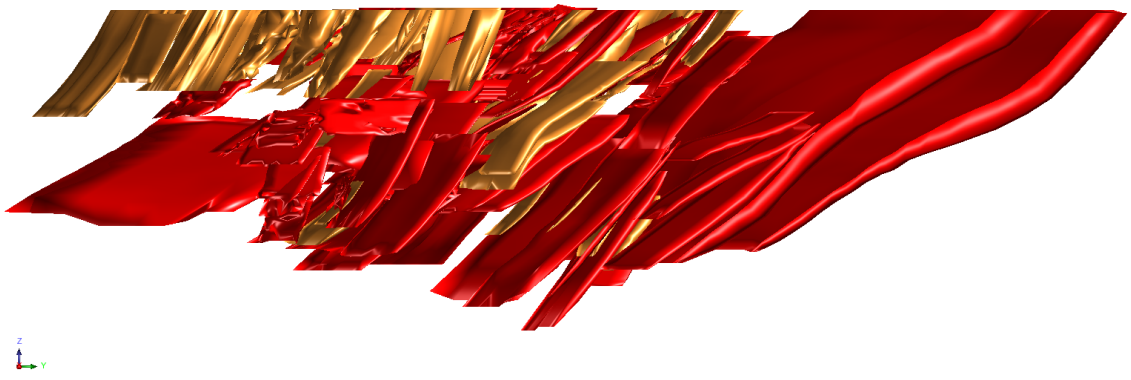

Figure 2.1: 3D–models of tonalitic-granodioritic-granitic gneisses (yellow) and pegmatitic granites (red). View from east.

Diabase dykes cut all the other rock types sharply with general direction in northeast–southwest direction. The width range from one centimetre to 2 metres. Modelled diabase dykes consist of seven individual units. It is hypothesized that diabase dykes could be found in en echelon formation which are supported by observations of spatially close dykes, but the findings are not used in the modelled units. [Aaltonen et al., 2010, chapter 6]

2.2.2 Ductile deformation

Deformation in Olkiluoto area has taken place in five phases. The first phase, D1, contains the oldest observed structures. The second phase, D2, is responsible for creating migmatites in the area with pervasive and homogeneous deformation. The third phase, D3 is divided to three main structural domains that are zones with pervasive foliation, zones with thrust related structural elements and zones with strike slip or oblique slip shear environments. The fourth phase, D4, is responsible for folding in north-northeast–south-southwest direction with dipping to east-southeast. The D4 structures are fragmented to the whole area where they are found as zones in north-northeast–south-southwest from width of a few metres to 500 metres. The fifth phase, D5, is distinguished as schistosity and crenulation in mica rich block and mica schists with medium grain size.

The goal of the spatial model for the ductile deformation is to identify tectonic units which can be described as statistically homogeneous based on their deformation properties. The area is divided to three independent tectonic units that are Northern Tectonic Unit, Central Tectonic Unit and Southern Tectonic Unit. Central Tectonic Unit and Southern Tectonic Unit each split to three subunits by deformation zones. The tectonic units have deformation zones bordering them. Five different deformation zones have been distinguished. The Northern Tectonic Unit is bordered by Selkänummi Deformation Zone. The Central Unit is cut by three deformation zones, which are Flutanperä Deformation Zone and two D_4 zones, D_4–1 and D_4–2. The Southern Unit has Liikla Shear Zone in the border with The Central Unit. D_4 zones continue to The Southern Unit from The Central Unit. The modelled 3D–structures are presented in Figure 2.2. [Aaltonen et al., 2010, chapter 7]

2.2.3 Alteration

Alteration section of the Geological model 2.0 consists of descriptions of alteration processes that are retrogressive metamorphism, hydrothermal alteration and surface weathering. The hydrothermal alteration products and hydrothermal geochemistry are described and their properties are characterized in the report. The main hydrothermal alteration minerals are clay minerals including illite, kaolinate, epidote, calcite, sulphide-quartz and quartz-sericite assemblages.

Spatial models of hydrothermal alteration consist of five 3D–models that include illitisation, kaolinisation, sulphidisation, carbonisation and sericitisation. The illitisation model describes rock volumes that are represented by 5 to 20 metres sections in drillcores. The strength of the rock is reduced by alteration.

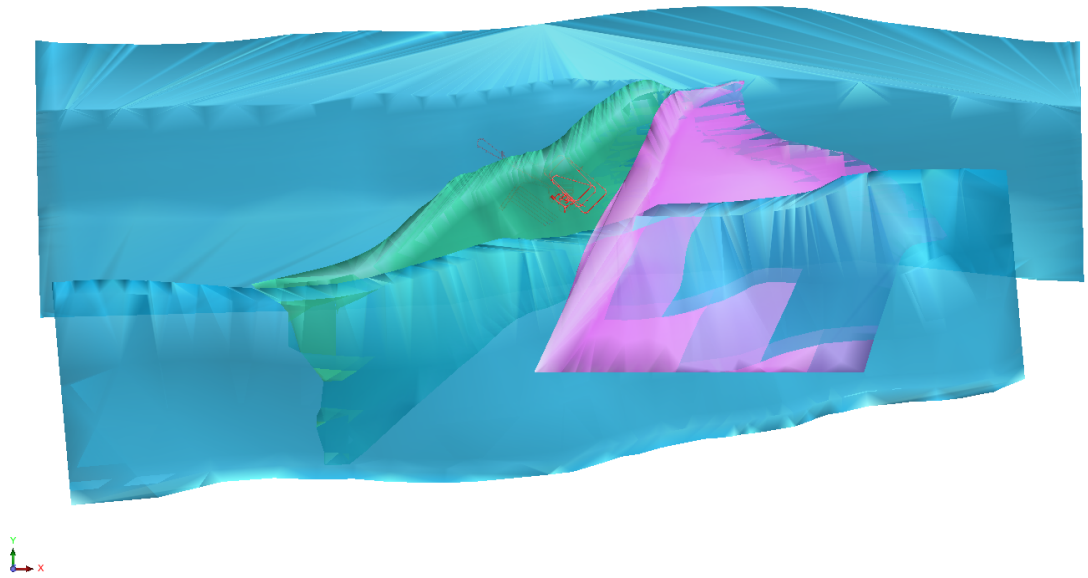

Figure 2.2: 3D–modelled ductile deformation zones excluding D_4-2, view from over the zones.

The kaolinisation model is divided to small lenses and spots that are commonly from centimetres to tens of metres thick in drillcore sections. Kaolinite forms 5 to 30 per cent of the rock volume. The modelled kaolinisited blocks describes rock volumes that are kaolisited but can contain also other alteration products.

The sulphidisation products are mainly pyrrhotite and also small amounts of pyritic coatings and pyrite stockworks exists. Pyrrhotite and graphite appear to occur together in graphite containing mica gneisses and migmatites. Pyrite is found in places with hydrothermal activity. The sulphide 3D–models delineate the main bedrock volumes that are sulphidicated, but small lenses can be found outside of the these units.

The carbonatisation products include calcite and dolomite. The amount of carbonates in the alterated volumes are considerable. Carbonates are commonly the main alteration products if it occurs and they are interpret as one of the latest alteration products. The unaltered rock typically does not include calcite in their volumes. The thicknesses of carbonates containing volumes range from a few metres to tens of metres.

The sericitisation commonly consist of hydrothermal alteration along thin bands. In drillcores the sericite is not found as pervasive. The fracture surfaces containing sericite are usually closed. Sericite is finely grained mass or coarse flakes up to one to two centimetres in diametre. [Aaltonen et al., 2010, chapter 8]

2.2.4 Brittle deformation model

Brittle deformation model considers geological features that are formed by brittle processes such as joints, veins and fissures. It also includes diabase dykes that are formed after the ductile deformation phases. It excludes veins filled with granitic, tonalitic, pegmatitic or amphibolitic material that are formed due to brittle deformation, but are now considered as intact rock. The fractures are found to have three separate fracture sets, which are fractures dipping gently between south to south-east, fractures dipping steeply to north-south and east-west directions, respectively. The first set follows the orientation of the ductile deformation. According to the investigations two thirds of the fractures contains filling minerals of which calcite, pyrite, kaolinite, chlorite and illite are the most common ones. The fracture surfaces are mostly irregularly shaped or semi-rough. Based on the fracture morphology almost 60 per cent of the fracture surfaces are curved wavy or undulating.

The brittle deformation model describes fracture kinematics and evolution of the movements. In brief the evolution can be divided to four phases where the first phase is includes shearing in D_3 and D_4 directions. The second phase contains diabase intrusions in association with extension in south-southeast–north-northwest direction. The third phase has extension event in east–northeast to northeast-southwest direction and the fourth phase includes contraction in north-south to north-east directions.

The brittle deformation model includes 3D–models of the deformation zones and single structural features. The one of the main motivation for these models is to examine possible routes of seismic events. The geological model 2.0 explains in detail about the criteria for deformation zones and single structural features. The spatial models are divided to two scales where the first considers site-scale brittle deformation zones and the second considers repository-scale brittle deformation zones. Criterion for the site-scale is that the lateral dimension of the zone exceeds 1 000 metres and it is verified by several drillholes or geophysical measurements. The repository scale considers zones that do not meet the criterion for site-scale zones. The confidences of the modelled zones are defined empirically. Typically the site-scale models have a better confidences compared to the repository-scale models due to higher amount of drillcore intersections. The updated 2.1 geological model in the Olkiluoto Site Description [Andersson et al., 2012] report that 3D brittle deformation zone models include 229 modelled fault zones of which 27 are labelled as site-scale and 202 are included in repository-scale. [Aaltonen et al., 2010, chapter 9]

2.2.5 Model comparison

The geological submodels are compared against each other and the correlations are reviewed to achieve consistent view. The Geological model 2.0 examines the relationship between the lithology with the ductile deformation and the brittle deformation, the ductile deformation with the brittle deformation, the alteration and the hydrogeology. [Aaltonen et al., 2010, chapter 10]

2.3 Stochastic Discrete Fracture Network

Discrete Fracture Network model, DFN, divides the bedrock volume to different fracture domains with their own statistical features. The model's source information includes fracture orientation, size, intensity, location, mineralogy, hydraulic and mechanical properties. The modelled submodels are expressed based on the theoretical probability distributions, which are inferred from the measured data. The DFN includes fracture orientation model, fracture spatial model and fracture intensity model. [Fox et al., 2012] & [Andersson et al., 2012]

2.4 Hydrogeological structure model

Latest hydrogeological structure model of the Olkiluoto is by Vaittinen et al. [2011]. The hydrogeological model is based on ductile deformation model, site-scale brittle deformation zones that are described in the geological model 2.0 [Aaltonen et al., 2010], hydrogeological and geophysical data gathered at Olkiluoto. The modelled hydrogeological structure model consist of 13 different hydrogeological zones that are HZ001, HZ008, HZ19A, HZ19B, HZ19C, HZ20A, HZ20B, HZ21, HZ21B, HZ039, HZ099, OL-BFZ100 and HZ146. Modelled zones are consided as planar or semi-planar and 3D–models comprises triangular nets without smoothing. The 3D–modelled hydrogeological zones are presented in Figure 2.3. [Vaittinen et al., 2011]

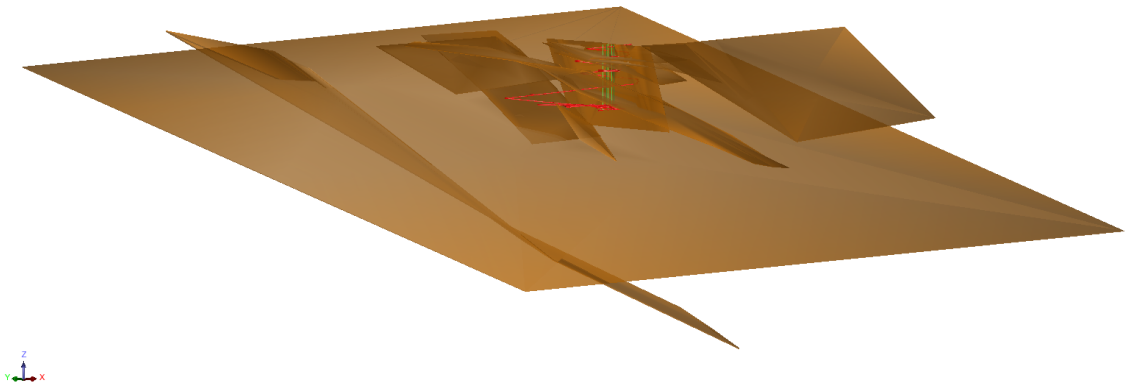

Figure 2.3: 3D–modelled hydrogeological structures in Olkiluoto, view from southwest.

2.5 Drillhole and –core investigations

Data acquired from the drillholes and -cores include geological, hydrogeological and geophysical measurements and investigations. The gathering is conducted from the POTTI-database with the permission of Posiva. The data are described in the Aaltonen et al. [2010] and Andersson et al. [2012] including their references. General information concerning drillholes is described in Table 2.1.

Table 2.1: General drillhole information

Data	Class	Type	Remarks
Hole collar	General	Point	Hole collar information includes collar coordinates and initial direction.
Hole survey	General	Point	Spatial information concerning relationship between hole depth and its coordinates.

Hydrogeological borehole interval data includes constant head injection tests and hydraulic conductivity. The data is gathered with Hydraulic Testing Unit, HTU and Posiva Flow Log, Difference flow method, PFL DIFF. The HTU method is described in Hämäläinen [2005] and PFL DIFF is introduced in Pöllänen [2009].

Hydrogeological borehole point data includes flow rate, fracture transmissivity and aperture, fracture transmissivity and head. The flow rate measurements are measured with PFL DIFF device that measures the amount of water that flow through sample space in specific time. The methods for flow rate measurements are described in Pöllänen [2009]. The hydrogeological data are described in Table 2.2.

Table 2.2: Hydrogeological measurements

Data	Class	Type	Units	Remarks
Constant head injection test	Hydrogeology	Interval	$m\,s^{-1}$	Measured with Hydraulic Testing Unit, HTU, hydraulic conductivity is reported for interval, calculated with the Moye-equation. [Hämäläinen, 2005]
Hydraulic conductivity	Hydrogeology	Inteval	$m\,s^{-1}$	Hydraulic conductivity measured for interval with Posiva Flow Log, Difference flow method, PFL DIFF. [Pöllänen, 2009]
Flow rate	Hydrogeology	Point	$ml\,h^{-1}$	Flow rate measured with PFL DIFF reported as point information. [Pöllänen, 2009]
Fracture transmissivity and aperture	Hydrogeology	Point	$m^2\,s^{-1}$, $ml\,h^{-1}$, m	Fracture transmissivity values, flow measurement and hydraulic aperture values for fractures.
Fracture transmissivity and head	Hydrogeology	Point	$m^2\,s^{-1}$, m, $ml\,s^{-1}$, m	Includes the transmissivity values, hydraulic head values, flow measurements with and without pumping and water head measurents with and without pumping. Measured with the PFL DIFF.

Geophysical borehole measurements include gamma-gamma density logging, magnetic susceptibility, resistivity logging, single point resistance, sonic full wave logging, spectral gamma logging and total gamma radiation. All the measurements are presented as point type data. The different measurement methods are described in detail in Posiva working reports Julkunen et al. [2004] and Majapuro [2006]. Corresponding geophysical borehole data are described in Table 2.3.

Table 2.3: Geophysical measurements

Data	Class	Type	Units	Remarks
Gamma-gamma density logging	Geophysics	Point	$g^3\,cm^{-1}$	Density values calculated with gamma-gamma method [Majapuro, 2006]
Magnetic susceptibility	Geophysics	Point	10^{-5}SI	Bedrock in situ magnetic susceptibility measurements. [Majapuro, 2006]
Resistivity logging	Geophysics	Point	$\Omega\,m$	Bedrock in situ normal resistivity measurements. [Majapuro, 2006]
Single point resistance	Geophysics	Point	Ω	Single point resistance measurements against drillhole wall. [Majapuro, 2006]
Sonic full wave logging	Geophysics	Point	$m\,s^{-1}$	Bedrock p- and s-wave velocities. Measured from borehole wall. [Majapuro, 2006]
Spectral gamma logging	Geophysics	Point	µR/h, ppm	Gamma radiation intensity and equivalent concentrations for uranium, thorium and potassium. [Ojala et al., 2007]
Total gamma radiation	Geophysics	Point	µR/h	Gamma radiation measurements from boreholes. [Ojala et al., 2007]

Geological and rock mechanical information includes interval type alteration mapping, ductile feature mapping, fracture frequency mapping, fractured zone mapping, lithological description, rock quality mapping, weathering degree and zone intersection. Point type geological information concerns fracture specific mapping data. The data are described in Table 2.4. [Aaltonen et al., 2010]

Table 2.4: Geological and rock mechanical measurements

Data	Class	Type	Remarks
Alteration mapping	Geology, Rock mechanics	Interval	Mineralogical alteration found in drillholes classified as fracture surface alteration or pervasive alteration.
Ductile feature mapping	Geology, Rock mechanics	Interval	Feature the element type, folitation type and structural direction of the ductile feature and rock type.
Fracture frequency	Geology, Rock mechanics	Interval	Includes the fracture frequency data for 1 metre intervals with natural and mechanically induced fractures separated and corresponding RQD-value for the interval.
Fracture zone mapping	Geology, Rock mechanics	Interval	Classification of the fracture zones based on the Finnish engineering geological classification system [Äikäs et al., 2000]
Lithological description	Geology, Rock mechanics	Interval	Includes rock type information and leucosome percentages [Mattila, 2006]
Rock quality mapping	Geology, Rock mechanics	Interval	Rock quality classification with Q-number parameters. [Äikäs et al., 2000]
Round mapping	Geology, Rock mechanics	Interval	Excavated tunnel mapping, round mapping. [Engström and Kemppainen, 2008]
Round mapping fracture	Geology, Rock mechanics	Interval	Excavated tunnel mapping, fracture information, round mapping. [Engström and Kemppainen, 2008]
Round mapping rock quality	Geology, Rock mechanics	Interval	Excavated tunnel mapping, rock quality, round mapping. [Engström and Kemppainen, 2008]
Round mapping Schmidt test	Geology, Rock mechanics	Interval	Excavated tunnel mapping, Schmidt hammer test, round mapping. [Engström and Kemppainen, 2008]
Systematic mapping fracture	Geology, Rock mechanics	Interval	Excavated tunnel mapping, fracture information, systematic mapping.[Engström and Kemppainen, 2008]
Systematic mapping fracture sets	Geology, Rock mechanics	Interval	Excavated tunnel mapping, fracture sets , systematic mapping.[Engström and Kemppainen, 2008]
Systematic mapping rock quality	Geology, Rock mechanics	Interval	Excavated tunnel mapping, rock quality, systematic mapping. [Engström and Kemppainen, 2008]
Weathering degree	Geology, Rock mechanics	Interval	Weathering classification. [Äikäs et al., 2000]
Zone intersection	Geology, Rock mechanics	Interval	Deformation zone intersection types and description. [Engström and Kemppainen, 2008]
Detailed fracture logging	Geology, Rock mechanics	Point	Detailed logging of fracture properties. [Engström and Kemppainen, 2008]

3 Theory

The aim of this section is to give tools for doing statistical analysis. The first part includes descriptive statistics that is used for general examination of the data. The descriptive statistics uses statistical parameters to describe the underlying distributions. The descriptive and summary terms are used synonymoysly in the following text. The statistical parameters are mean values, variance, covariance and quartiles with minimum and maximum values. The first part also gives short introduction to distribution that are common in geosciences.

The second part of this section handles the different test procedures and the applied statistical tools. These include Kolmogorov-Smirnov test for equality, Shapiro-Wilks test for Normality and one sample t-test. The second part also includes procedures to perform principal factor analysis using principal component analysis as an intermediate step, doing multiple regression analysis using the stepwise backward model selection with the Akaike information criteria.

3.1 Descriptive statistics

The descriptive statistics are calculated for sample with an assumption that the sample $X = \{x_1, x_2, \cdots, x_n\}$ has been taken randomly from the parent population with specific distribution and the samples are independent of each other. Sample size n refers to the number of random observations in a given sample. [Kreyszig, 2006]

Arithmetic mean ${}^{A}\overline{x}$, described in equation 1, is calculated as average of the sample.

$$ {}^{A}\overline{x} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1} $$

where n is the sample size of x. Geometric mean ${}^{G}\overline{x}$ is calculated as the nth root of the sample product. The sample values need to be positive real numbers.

$$ {}^{G}\overline{x} = \left(\prod_{i=1}^{n} x_i\right)^{1/n} \tag{2} $$

where n is the sample size. The geometric mean can transformed in to the logarithm form where the geometric mean is the antilogarithm of the arithmetic mean of the logarithmic values.

$$ {}^{G}\overline{x} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \ln x_i\right) \tag{3} $$

Sample variance s^2 is calculated as a sum of squared difference between sample mean and observations.

$$ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{4} $$

Covariance between two samples s_{hj}, where h and j are samples from multivariate data, is calculated with the help of samples arithmetic means. In the case of $h = j$ sample covariance reduces to unbiased sample variance.

$$s_{hj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ih} - \mu_h)(x_{ij} - \mu_j) \quad \text{for } h, j = 1, \ldots, \mathrm{k} \tag{5}$$

Randomly sampled values follow the parent distribution, where the distributions in geosciences are commonly normal, lognormal or Poisson [Davis, 2002, chapter 2]. Many natural phenomena follow the normal distribution. The normal distribution is described with mean value of μ and variance σ^2. The probability distribution function for normal distribution is presented in equation 6.

$$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6}$$

In order to be able to calculate normal distributions, the mean and the variance of the population needs to be known or estimated. The population mean μ is estimated with arithmetic sample mean $\overline{x}$ and population variance σ^2 is estimated with sample variance s^2. [Kreyszig, 2006]

Sample follows lognormal distribution if its logarithmic transformation follows normal distribution. The transformation can be done with direct conversion of values if all the sample values are higher than zero. If sample includes zero values a box-cox transformation, with $\lambda_1 = 0$ [Box and Cox, 1964], can be used which is described in equation 7.

$$x_i^{(\lambda)} = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \\ \log(y + \lambda_2) & \text{if } \lambda_1 = 0 \end{cases} \tag{7}$$

The Poisson distribution is encountered when the data can described as count data. The Poisson distribution can be transformed with square root transformation if the data are highly skewed. [Davis, 2002, chapter 2]

Sample skewness describes how the sample values are distributed around the mean value. Positive skewness value means that the distribution has long tail to the higher values than mean with most of the values being less than the mean value and negative skewness values indicates the opposite. The normal distribution has zero skewness. The skewness function is produced with the help of central moment function described in equation 8

$$m_k = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^{\ k} \tag{8}$$

The skewness in equation 9 is presented as a function of second and third central moments of the sample. [Davis, 2002]

$$\text{skew} = \frac{m_3}{{m_2}^{\frac{3}{2}}} \tag{9}$$

The sample kurtosis describes the peakedness of the distribution. The normal distribution's kurtosis is one. The higher values than one indicate that the sample has more peakedness than normal distribution and lower value indicates that the sample values are more spread. The kurtosis is calculated as a function of fourth and second central moments as is presented in equation 10. [Davis, 2002]

$$\text{kurt} = \frac{m_4}{{m_2}^2} \tag{10}$$

Sample quantiles describes the cumulative distribution. q-quantiles divide data to q equal groups and when the $q = 4$ quantiles are called quartiles. The first quartile, Q_1, is the value that divides the data to lowest 25% and highest 75%. Second quartile, Q_2, is synonymous to median, Me and it divides the data to lowest 50% and highest 50%. Third quartile, Q_3, divides the data to lowest 75% and highest 25%. In this thesis the quartiles are calculated with *type* 7 that is described in Hyndman and Fan [1996].

$$Q_t = (1 - \gamma)x_j + \gamma x_{j+1} \tag{11}$$

where γ is function of j and m described in equation 12. j value is the jth order statistic of sample, described by equation 13 and m is function of p, described in equation 14.

$$\gamma = np + m - j \tag{12}$$

where j is jth order statistic, n is sample size and p as described in equation 15 and m is the function of p, described in equation 14.

$$j = \lfloor np + m \rfloor \tag{13}$$

where flooring function $\lfloor y \rfloor$ is the largest integer not greater than y.

$$m = 1 - p \tag{14}$$

where p is described in the equation 15.

$$p = \frac{t}{q} \tag{15}$$

where t is tth quantile and q is the chosen quantile number. [Hyndman and Fan, 1996]

3.2 Two-sample Kolmogorov-Smirnov test for equality

Two-sample Kolmogorov-Smirnov test is nonparametric test that is used to test if two different samples are sampled from the same distribution. The null hypothesis is that the two samples follow the same distribution.

The empirical cumulative distribution function is

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{x_i<x} \tag{16}$$

where $I_{x_i<x}$ is indicator function with values 1 if the sample value x_i is smaller than test value x and 0 otherwise. The Kolmogorov-Smirnov test statistic is calculated with empirical distributions of each sample. Test statistic is the largest absolute difference between two empirical distributions.

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{17}$$

where the $\sup_x |y(x)|$ is supremum of y, which means the least upper bound of the absolute value of y.

The null hypothesis is rejected at level α if test statistic is larger than $D_{critical}$.

$$D_{critical} = c(\alpha)\sqrt{\frac{n+m}{nm}} \tag{18}$$

where $c(\alpha)$ is constant based on the critical level. Values for $c(\alpha)$ can be found in statistical literature. In this thesis $\alpha = 0.05$ is used and the corresponding value is $c(0.05) = 1.36$. [Conover, 1999]

3.3 Shapiro-Wilks test for Normality

Shapiro-Wilks test for Normality is for testing if the sample comes from the normal distribution. The test statistic W_{S-W} is compared to predetermined level of significance α. The null hypothesis assumes that the sample has a normal distribution and if the test value is under the threshold level, null hypothesis is rejected. The test statistics is calculated as

$$W_{S-W} = \frac{(\sum_{i=1}^{n} a_i \cdot x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{19}$$

where $x_{(i)}$ is the i-th order statistic of the value x_i, $\overline{x}$ is the mean value of the sample and a_i is constant that is calculated based on the equation 20.

$$(a_1, \ldots, a_n) = \frac{\mathbf{m}^\intercal \mathbf{Q}^{-1}}{(\mathbf{m}^\intercal \mathbf{Q}^{-1}\mathbf{Q}^{-1}\mathbf{m})^{\frac{1}{2}}} \tag{20}$$

where m_i is the expected value of the i:th order statistics sampled from the standard normal distribution. $\mathbf{Q}$ is the variance-covariance matrix of the expected values.

On large sample sizes the test may give statistically significant results with slight deviation from normal distribution, which is the reason for to use Quantile-Quantile plot, Q-Q plot, to confirm the results. [Davis, 2002]

3.4 Principal components analysis and Principal factor analysis

Principle component analysis, PCA, is data reduction procedure that keeps data variance as maximum as possible.The PCA is calculated from the standardised datasets as the eigenvalues and eigenvectors of the variance-covariance matrix Q that is the same as the correlation matrix R in standardised case. The standardisation procedure ensures that all the variables have same weigth for the analysis.

$$\mathbf{Q}_{nk}^{std} = \begin{bmatrix} {s_1}^2 & s_{12} & \cdots & s_{1k} \\ s_{21} & {s_2}^2 & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & {s_k}^2 \end{bmatrix}$$

Eigenvalues are calculated by solving the following equation for λ.

$$\mathbf{Qx} = \lambda\mathbf{x} \tag{21}$$

Equation 21 can be rewritten by moving the right-handside of the equation to the left-handside as is expressed in equation 22.

$$(\mathbf{Q} - \lambda\mathbf{I})\mathbf{x} = 0 \tag{22}$$

In the case of non-trivial solutions, where trivial solution is $x_{ij} = 0$ for all the i and j, it follows that the determinant of the $\mathbf{Q} - \lambda\mathbf{I}$ equals zero, based on the Cramer's theory. [Kreyszig, 2006, chapter 7]

$$|\mathbf{Q} - \lambda\mathbf{I}| = 0 \tag{23}$$

In the case that all the rows in the variance-covariance matrix $\mathbf{Q}$ are independent the rank equals matrix order. If some of the rows are dependent then the rank of the variance-covariance matrix $\mathbf{Q}$ is smaller than order of the independent variance-covariance matrix $\mathbf{Q}$. The variance-covariance matrix $\mathbf{Q}$ being symmetrical means that all the eigenvalues are real. For each eigenvalues λ_j the corresponding eigenvectors are calculated by solving the $\mathbf{x}$ from the eigenvalues from equation 22.

The factor analysis based on the principal component analysis is called principal factor analysis. Eigenvectors are converted to principal factors with the square roots of eigenvalues. In matrix presentation the principal factor matrix $\mathbf{A}$ is calculated by multiplying eigenvector $\mathbf{U}$ with vector $\mathbf{\Lambda}$ where $\mathbf{\Lambda}$ is diagonal matrix with the eigenvalues on the diagonal in descending order. The elements in the $\mathbf{A}$ matrix are called principal factor loadings and they describe how different variables relate to each other in each factor.

$$\mathbf{A} = \mathbf{U\Lambda} \tag{24}$$

The principal factor analysis on k × k matrix consist of k-factors and by calculating the relative proportion of each eigenvalues one can find the amount that each of the eigenvalues contributes for variance in the principal factors. The amount of principal

factors can be reduced by rejecting the least significant ones depending on the goal of the principal factor analysis.

$$\%_{\text{trace}}^{\lambda_i} = \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j} \quad \text{for } i = 1, \ldots, \mathrm{k} \tag{25}$$

The cumulative percentage from the highest percentage components to lowest ones describes the amount that cumulative collection of the components explains the variance in the data. The amount final principal factors explain the variance in the data depends on the criterion used for choosing the number of principal factors. One popular method is to use the Kaiser criterion, which means that eigenvalues less than 1 are excluded from the analysis. [Davis, 2002]

3.5 Multiple linear regression analysis

In multiple linear regression, a linear model of the dependent variable is constructed based on the independent variables and residual, where $\varepsilon_j \sim N(0, {\sigma_\varepsilon}^2)$.

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk} + \varepsilon_j \quad \text{for } j = 1, \ldots, \mathrm{n} \tag{26}$$

Multiple linear regression model can be expressed in the matrix form, where $\mathbf{y}$ is the size of $\mathrm{n} \times 1$, $\mathbf{X}$ size is $\mathrm{n} \times (\mathrm{k}+1)$ and $\boldsymbol{\beta}$ have the size of $(\mathrm{k}+1) \times 1$ and $\boldsymbol{\varepsilon}$ have size $\mathrm{n} \times 1$. The $\mathbf{X}$ matrix has an added first column of ones describing the constant that is linked for β_0 term.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{27}$$

Based on the Gauss-Markov theorem, the fitting of the model is done by minimizing the sum of squared residuals. This is called the Ordinary Least-Squares Estimation method, described in equation 28.[Plackett, 1950]

$$\min \sum_{j=1}^{\mathrm{n}} \widehat{\varepsilon}_j^{\,2} = \min \sum_{j=1}^{n} (y_j - b_0 - b_1 \cdot x_{j1} - b_2 \cdot x_{j2} - \cdots - b_k \cdot x_{jk})^2 \tag{28}$$

where b_j is the estimated coefficient corresponding to coefficient β_j. The coefficient estimates b_j are calculated as described in equation 29. [Davis, 2002]

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y} \tag{29}$$

3.5.1 Multiple polynomial regression analysis

Multiple polynomial regression is special case of the multiple linear regression where the dependent variable is modelled as nth order polynomial. The regression procedure is same as the multiple linear regression. Depending on the modeller interest nth order variables can be nth order combination of the independent variables.

$$\begin{aligned} y_j = \beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk} + \cdots + \sum \beta_t ({x_w}^u \, {x_v}^o) + \varepsilon_j \\ \text{for } j = 1, \ldots, \mathrm{n} \end{aligned} \tag{30}$$

where $1 \leq (\mathrm{v}, \mathrm{w}) \leq n$ and $(\mathrm{u}, \mathrm{o}) \leq n$th order. In the case of $\mathrm{u} = 0$ and $\mathrm{o} = 0$ polynomial regression model changes back to multiple linear regression when β_t coeffient is merged in to the β_0.

3.6 One-Way Analysis of Variance

The regression model results can be tested with One-Way Analysis of Variance. Analysis of variance, ANOVA, is used to test if the sample means for all of the variable groups are same. ANOVA assumes that sample X_j is a random and independent sample from the normal population with the mean μ_j and variance ${\sigma_j}^2$.

The null hypothesis, H_0, assumes that the mean for each group is equal to each other. The alternative hypothesis, H_1, is that at least two of the means are statistically different.

The data in ANOVA are divided to groups or populations. The groups and total means are calculated for the dataset. The sum of squares for groups and the total are calculated with the help of groups and total mean respectively.

$$SSgroup_j = \sum_{i=1}^{n_j}(x_{ij} - \overline{x}_j)^2 \tag{31}$$

$$SS_{total} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{x})^2 \tag{32}$$

Sum of squares for error is solved from the equation for total sum of squares

$$SS_{total} = \sum_{j=1}^{k}(SSgroup_j) + SS_{error} \tag{33}$$

The mean sum of squares are calculated based on the degrees of freedom in the sum of squares.

$$MSgroup_j = \frac{SSgroup_j}{(n_j - 1)} \tag{34}$$

$$MS_{error} = \frac{SS_{error}}{\sum_{j=1}^{k}(n_j) \times k - \sum_{j=1}^{k}(n_j)} \tag{35}$$

The F-test value is calculated from the ratio between the mean sum of squares. It follows F-distribution with the parameters (n-1,nk-n). Large F-test values lead to rejecting the null hypothesis.

$$\mathrm{F}_j = \frac{MSgroup_j}{MS_{error}} \tag{36}$$

The p-value is the smallest level of significance at which the null hypothesis would be rejected for a specific test value. Depending on the chosen value of the level of significance the null hypothesis is either kept or rejected.

In the regression analysis ANOVA is used to test if regression coefficients equal to zero. The degrees of freedom for coefficients are 1 instead of $n_j - 1$ as it would be in normal case.

The results of the ANOVA are summarized in the ANOVA table. [Davis, 2002]

Table 3.1: Analysis of variance table

Sum of Squares	Degrees of Freedom	Mean Square	F-value	p-value
$SSgroup_1$	n_1-1	$MSgroup_1$	F_1	p_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$SSgroup_j$	n_j-1	$MSgroup_j$	F_j	p_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$SSgroup_k$	n_k-1	$MSgroup_k$	F_k	p_k
SS_{error}	$\sum_{j=1}^{k}(n) \times k - \sum_{j=1}^{k}(n)$	MS_{error}		
SS_{total}	$\sum_{j=1}^{k}(n) - 1$			

3.7 Stepwise backward model selection

Stepwise backward regression starts with the regression model using all the independent variables and subsequently, a step-by-step algorithm to remove iteratively the least significant independent variables one for each round. Iteration stops after all the variables are significant. The significance is tested against either a predetermined p-value, based on for example the t-test or z-test, or the information criterion such as Akaike information criterion, AIC, or Bayesian information criterion, BIC.

3.8 Akaike information criterion

The significance of model paratemeter can be tested against the Akaike information criterion, AIC. It compares the goodness of fit with the complexity of the model. Akaike information criterion is calculated with the number of parameters, k_j and maximum likelihood value L_j.

$$AIC = 2k - 2\ln L_j \tag{37}$$

where k is the number of parameters in model, L_j is the maximum value for the likelihood function and ln refers to natural logarithm. The information criterion can be presented in another form when comparing different linear regression models. Then the AIC is presented with the residual sum of squares.

$$AIC = n\ln(\frac{RSS}{n}) + 2k \tag{38}$$

where RSS is residual sum of squares, $\sum {\varepsilon_j}^2$ and n is the sample size. According to Burnham and Anderson [2004] the ratio of the n to k is small should one be using the Akaike information criteria with correction, AIC_C. This extended version includes a correction term for the finite sample size. In the case that the samples have equal number of observations the AIC and AIC_C give the same relative results, which means that the AIC can be used instead of AIC_C. [Burnham and Anderson, 2004]

$$AIC_C = AIC + \frac{2k(k+1)}{n-k-1} \tag{39}$$

Bayesian information criterian can be derived from the AIC by changing the k's coefficient from 2 to $\ln(n) + \ln(2\pi)$ and can be approximated by $\ln(n)$ when n is large.

$$BIC = k[\ln(n) + \ln(2\pi)] - 2\ln(L_j) \tag{40}$$

where k is number of parameters, n is the sample size and L_j is the maximum likelihood estimator. [Johnson and Wichern, 2007]

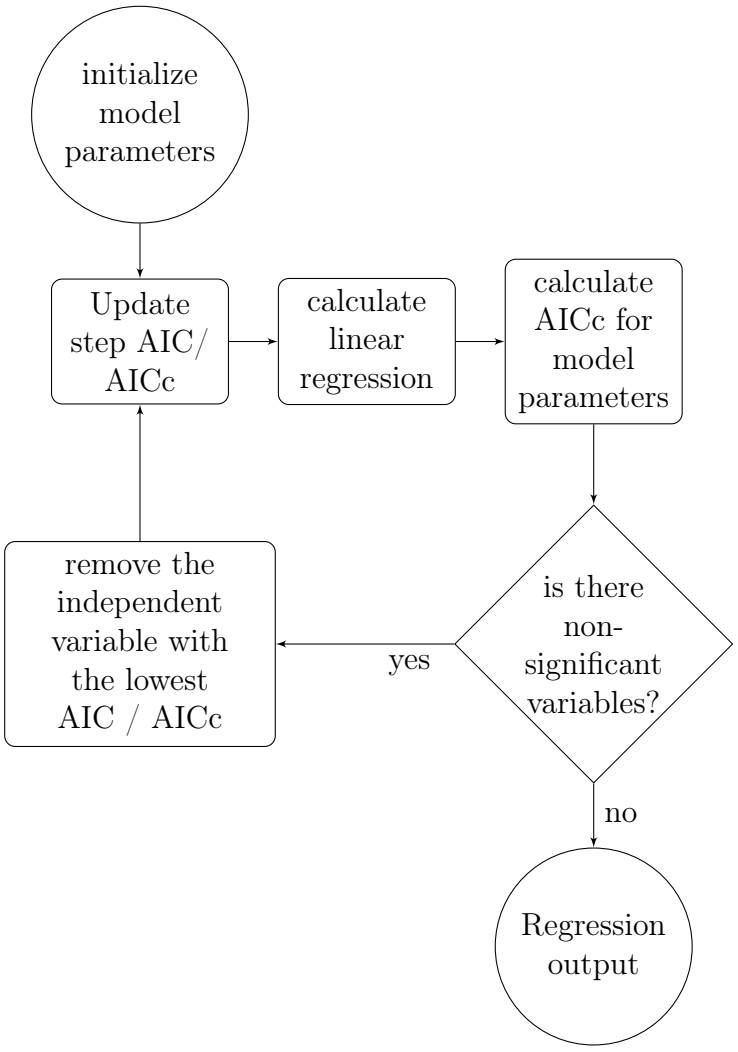

Figure 3.1: Stepwise multiple linear regression procedure with Akaike information criteria

3.9 One sample t-test

To test models parameters significances against null hypotheses one can use t-tests. Test statistic is from the Student's t-distribution with $(n - 2)$ degrees of freedom. Student's t-test assumes that the distribution of the population means have normal

distribution.

$$t = \frac{\widehat{\beta}_j - \theta_j}{SE_{\widehat{\beta}_j}} \quad \sim t(\mathrm{n}-2) \tag{41}$$

where $\widehat{\beta}_j$ is maximum likelihood estimate for parameter, θ_j is the assumed value specified null hypothesis, $SE_{\widehat{\beta}_j}$ is standard error of the estimated parameter and n sample size. The standard error of the estimated parameter is calculated from the sum of squares of difference between the observed value of dependent variable and estimated value of the dependent variable that are divided by the sum of squares of difference between the observed value of independent variable j and its mean value.

$$SE_{\widehat{\beta}_j} = \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}} \tag{42}$$

where y_i is the ith sample from dependent variable, $\widehat{y}$ is the estimated value from regression model, x_{ij} is the ith sample of the independent variable j and $\overline{x}_j$ is the mean the jth independent variable. The sum of squares of difference in dependent variable in equation 42 can be presented as residual sum of squares, RSS. [Davis, 2002, chapter 2]

$$SE_{\widehat{\beta}_j} = \frac{\sqrt{\frac{1}{n-2}RSS}}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}} \tag{43}$$

4 Methods and testing

4.1 Sample space

Sample space for statistical analyses is selected from circular section with 4 kilometre radius that has the middle point in 6 792 000 N 1 526 000 E (KKJ1–coordinate system) and depth between 300–500 metres. The depth interval minimizes the influence of the more fractured surface bedrock and focus on the bedrock properties at the final disposal levels. This space has 44 OL-KR drillholes with the transmissivity data measured in 182 spots distributed between 26 holes. The combined length of the drillholes within the studied volume is 7 911 metres whereas the total length of the drillholes in Olkiluoto is 34 013 metres.

4.2 Sample data

The objective of the analysis is to study the relationship between bedrock properties and the transmissivity. The property selections are based on the potential relationship with the transmissivity and are as well in the interest of the modeller. These properties are fracture frequency, single-point resistance, lithological rock types and joint surface properties including joint surface type, carbonate, sulphide and clay contents.

The sample data are first described with its different statistics and are also presented in the summary tables and after that the describing figures are added for each parameter. The figures include combined histogram and kernel density plot that describes parameters distribution. The second type of figure is the Box-Whisker plot that also describes parameters distribution. Box-Whisker plot is used when categorical data are visualised based on the different parameter. Third figure type is Quantile-Quantile plot that is a scatter plot of the parameter's measurements against theoretical distribution.

The histograms are created by dividing data to appropriate bins where the corresponding area of the bin describes the weight that the bin has. In the case of equal size bins the height of the bins describes the counts or density of the samples in the bin. Kernel density plots are created by replacing every datum point with appropriate distribution that usually is normal distribution and by summing the created distributions one obtains continuous distribution of the sample. This procedure enables to approximate the discrete data in continuable way.

Vertical Box-Whisker plot is created by drawing the median of the sample to the plot as a horizontal line. The box containing median is drawed between the values in the sample that are the smallest and largest compared to the first and third quartiles respectively. The lines leaving the box are called whiskers and in the upper case extend to the highest value that is within $1.5 \times IQR$ of the third quartile, where IQR is meaning the inter-quartile range. In the lower case the whisker goes to the lowest value that is within $1.5 \times IQR$ of the first quartile.

$$IQR = Q_3 - Q_1 \tag{44}$$

where IQR is inter-quartile range, Q_3 is third quartile and Q_1 first quartile. Data that are not between whiskers are marked with points and can be considered as outliars. Notches are drawn from the median to $(1.58 \times IQR)/\sqrt{(n)}$ where the n is the sample size. The notches give roughly estimates of 95% confidence intervals for median. If two notches do not overlap it is an indication that the two medians are not the same. In the case that notch goes farther from the median than Q_1 or Q_3, may indicate low confidence for the boxplot. [McGill et al., 1978]

Quantile-Quantile plot is used to compare sample's distribution to the theoretical distribution. Sample data are sorted in ascending order and their ranks are calculated. Theoretical quantiles are calculated from the inverse cumulative distribution function based on the sample ranks. Scatterplot based on the theoretical quantiles and sample data are plotted against each other. If the theoretical distribution corresponds to the sample's distribution, scatter points should create a straight line. Visual help is made by drawing a straight line through the mean value with a slope of standard deviation.

A transmissivity value expresses the capability of water to flow through fractures or fracture zones. The measurements are reported as metre per second $\mathrm{m^2\,s^{-1}}$. The count of the discrete transmissivity measurement in the sample data are 372 where the mean is $2.01 \times 10^{-}6\ \mathrm{m^2\,s^{-1}}$, its median $2.41 \times 10^{-8}\ \mathrm{m^2\,s^{-1}}$ and variance $7.62 \times 10^{-11}\ \mathrm{m^2\,s^{-1}}$. The cut-out value of the measurement probe is dependent of the measurement conditions and are described in Aaltonen et al. [2010]. This means that there may exists lower transmissivity values than that are encountered in the data due to the cut-out value that is commonly $10^{-9}\ \mathrm{m^2\,s^{-1}}$ [Ahokas et al., 2012]. The maximum value was $10^{-5}\ \mathrm{m^2\,s^{-1}}$, which is four decades larger than the cut out value. Depending on the drillhole the transmissivity values have been measured between 1 to 9 times.

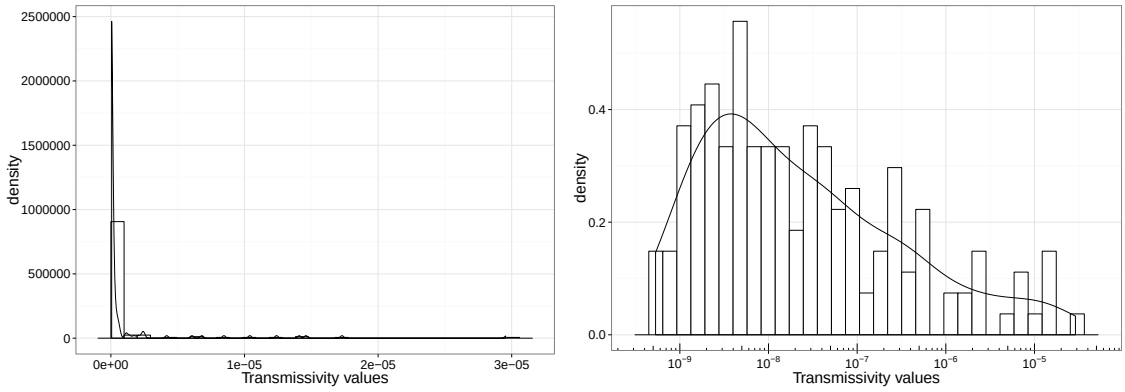

Figure 4.1: Transmissivity $[\mathrm{m^2\,s^{-1}}]$ histogram and kernel density plots in original (left) and logarithmic in base 10 (right) cases.

Taking the maximum measurement in each place reduces the number of measurements. This procedure reduces the amount of the data to 179 individual measurements. After excluding measurements less than the cut-out value the amount

decreases to 176 measurements. The corresponding summary statistics is described in Table 4.1. Based on the visual inspection of the Q-Q-plot an assumption is made that data are roughly logarithmically distributed.

Table 4.1: Summary: Transmissivity - All samples

$^{\dagger}\mu$ [$m^2 s^{-1}$]			s^2 [m^4/h^2]	s [$m^2 s^{-1}$]	$^{\dagger\dagger}skew$	$^{\dagger\dagger}kurt$
$2.05\times10^{-6}/3.14\times10^{-8}$			1.18×10^{-10}	1.09×10^{-5}	9.53/0.79	105.24/2.94
min [$m^2 s^{-1}$]	Q_1 [$m^2 s^{-1}$]	Me [$m^2 s^{-1}$]	Q_3 [$m^2 s^{-1}$]	max [$m^2 s^{-1}$]	IQR [$m^2 s^{-1}$]	n
4.64×10^{-10}	3.21×10^{-9}	2.11×10^{-8}	1.89×10^{-7}	1.28×10^{-4}	1.85×10^{-7}	176

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, IQR = interquartile range, n= sample size

† arithmetic mean / geometric mean

†† $y/y_{log_{10}}$

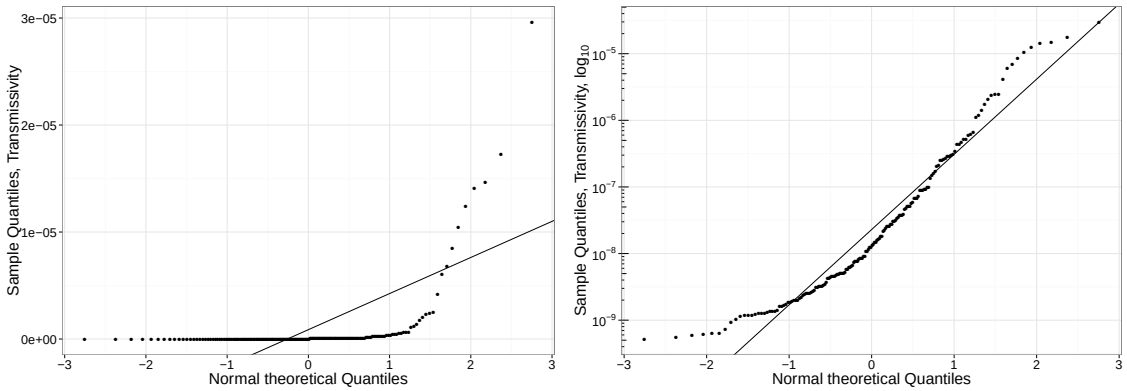

Figure 4.2: Quantile-Quantile plots of transmissivity [$m^2 s^{-1}$] in original (left) and logarithmic in base 10 (right) cases.

The fracture frequency data describes the amount of fractures for given interval. Dividing the fracture count with interval length fracture frequency can be expressed as fractures per metre [m^{-1}]. Sample data has 6628 metres of measurements where the mean value for the fracture frequency is 1.90 m^{-1} and the median is 1 m^{-1}. The variance of the data is 9.32 m^{-1} with the minimum value being zero, which

mean that bedrock does not have any fractures on that interval. Maximum value is reported as 28 m^{-1}. The corresponding summary statistics is described in Table 4.2. Based on the way fracture frequency data are gathered an assumption can be made that fracture frequency data has a Poisson distribution. Kernel density plot and histogram support the assumption.

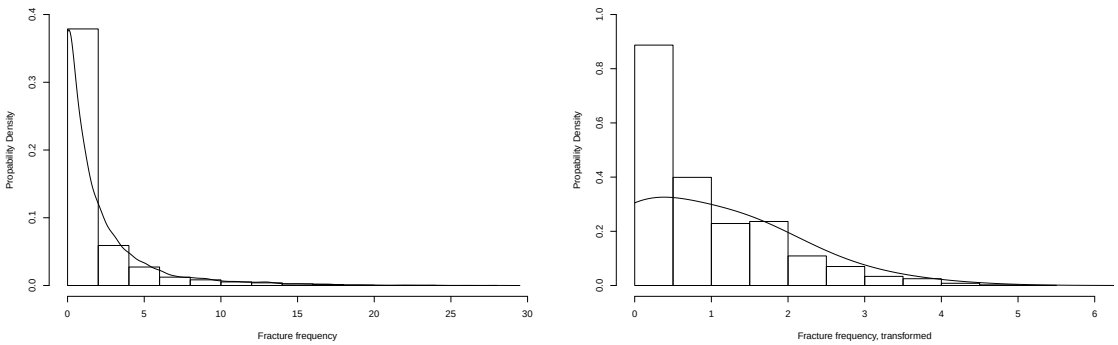

Figure 4.3: Fracture frequency [m^{-1}] histogram and kernel density plot (left) and histogram and kernel density plot with transformed data(right).

Table 4.2: Summary: Fracture frequency - All samples

$^{\dagger}\mu$ [m^{-1}]			s^2[m^{-2}]	s [m^{-1}]	*skew*	*kurt*
1.87/$^{\dagger\dagger}NaN$			8.99	3.00	2.82	13.33
min [m^{-1}]	Q_1 [m^{-1}]	Me [m^{-1}]	Q_3 [m^{-1}]	max [m^{-1}]	IQR [m^{-1}]	n
0	0	1	2	28	2	8560

s^2=sample variance, s=corrected sample standard deviation, *skew*=skewness, *kurt*=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, IQR = interquartile range, n= sample size

† arithmetic mean / geometric mean

$^{\dagger\dagger}NaN$ = Not a Number

Single point resistance (SPR) values describe the electrical resistance [ω] of drill-hole's sidewall. The sample data includes 213074 single point resistance measurements with the mean value of 1988.26 Ω m and median being 1419.80 Ω m. The

minimum resistance value is 0.01 Ω m, maximum value is 79657.80 Ω m and the sample variance is 4787566.25 $\Omega^2\,m^2$.

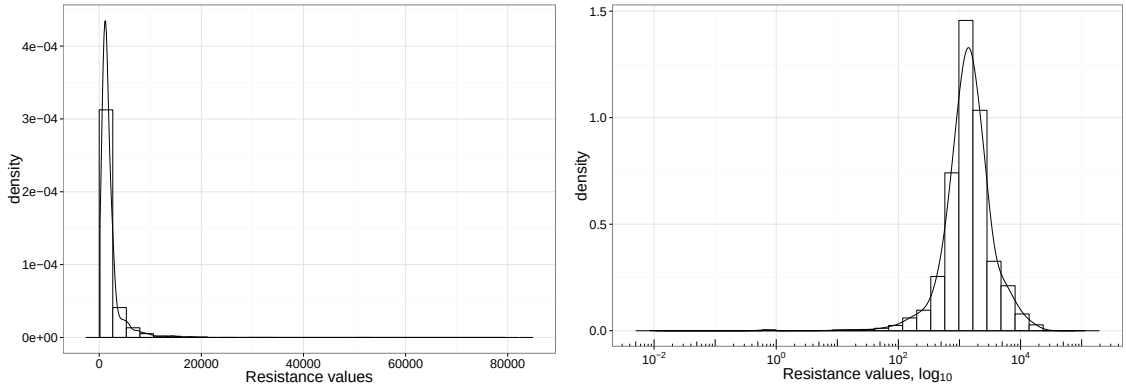

Figure 4.4: Resistance [Ω m] histogram and kernel density plots in original (left) and logarithmic in base 10 (right) cases.

The corresponding descriptive statistics is described in Table 4.3. Based on the visual inspection of Q-Q plot and an assumption is made that data are roughly logarithmically distributed.

Table 4.3: Summary: Single point resistance - All samples

$^{\dagger}\mu$ [Ω m]			s^2 [$\Omega^2\,m^2$]	s [Ω m]	$^{\dagger\dagger}skew$	$^{\dagger\dagger}kurt$
1988.26/1378.97			4.79×10^6	2188.05	4.94/−1.53	59.76/15.13
min [Ω m]	Q_1 [Ω m]	Me [Ω m]	Q_3 [Ω m]	max [Ω m]	IQR [Ω m]	n
0.01	908.03	1419.81	2250.09	7.97×10^4	1342.06	213074

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, IQR = interquartile range, n= sample size

† arithmetic mean / geometric mean

†† $y/y_{log_{10}}$

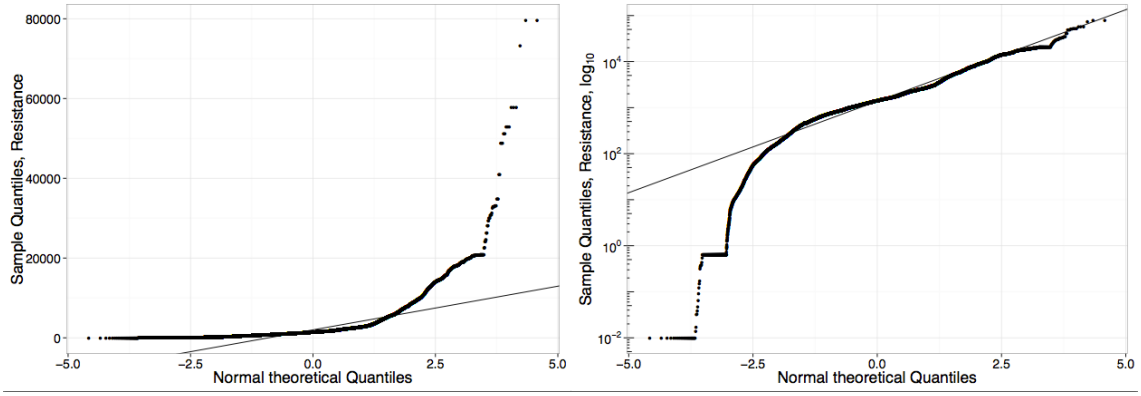

Figure 4.5: Quantile-Quantile plots of single point resistance [Ω m] in original (left) and logarithmic in base 10 (right) cases.

The lithological description of rock types of the Olkiluoto consists of eleven distinguishable types. The total length of the rock type data is 11047.33 metres and the total occurrences of the rock types are 1180. The main rock types in the sample data are diatexitic gneiss (DGN), with 2351.87 metres and 253 occurrences, pegmatitic granite (PGR), with 1973.39 metres and 360 occurrences, and veined gneiss (VGN), with 5198.30 metres and 377 occurrences. Other noticeable rock types are mica gneiss (MGN), with 705.15 metres and 100 occurrences, tonalitic-granodioritic-granitic gneiss (TGG), with 553.59 metres and 47 occurrences, K-feldspar porphyry (KFP), with 92.81 metres and 18 occurrences and mafic gneiss (MFGN), with 86.31 metres and 23 occurrences.

Table 4.4: Summary: Lithology - All samples

Lithological unit†	Abbreviation	Occurrence	Total length [m]
Veined gneiss	VGN	377	5198.30
Diatexitic gneiss	DGN	235	2351.87
Pegmatitic granite	PGR	360	1973.39
Mica gneiss	MGN	100	705.15
Tonalitic-granodioritic-granitic gneiss	TGG	47	553.59
K-feldspar porphyry	KFP	18	92.81
Mafic gneiss	MFGN	23	86.31
Quartz gneiss	QGN	11	45.95
Stromatic gneiss	SGN	5	35.61
Diabase	MDB	2	3.20
Crushed rock	CRUSH	2	1.15

† Abbreviation and explanation from [Aaltonen et al., 2010]

Joint surface properties used in analysis include joint surface type and joint surface filling. The total amount of the fractures in the sample area included 13 977 fractures within 7590.75 metres. The joint surface type is divided to two different classes. The first one is a general geological fracture type that includes eight different subgroups from which eight of them were founded in the sample space.

The description and summary of the occurrences are presented in Table 4.5. The second class is a geological fracture type that includes eight different groups from which six plus null groups where found at the sample volume. Geological fracture types are excluded from the grouping and regression analysis.

Table 4.5: Joint surface, general geological fracture types - All samples

Surface type†	Abbreviation	Occurrence
Single fracture	sf	8536
Hair dyke	hd	2506
Single shear	ss	1996
Fracture zone, unweathered	fz	407
Paleofracture	pf	333
Shear zone	sz	94
Break	kt	74
Fracture zone, weathered	fw	30
NULL	other	1

† Abbreviation and explanation from [Tammisto and Palmén, 2011]

Table 4.6: Joint surface, geological fracture types- All samples

Surface type††	Abbreviation	Occurrence
Curved	c	4835
Breakage	m	2706
Unknown infill	q	2273
Slickenside	h	2088
Planar	t	783
Dyke or paleofracture	j	655
Weathered fracture	w	395
Fracture filled with soft material	f	109
NULL	other	133

†† Abbreviation and explanation from [Tammisto and Palmén, 2011]

The sample data in the joint surface filling are focusing on the carbonate, sulphide

and clay contents. The values in the sample are the product of the joint filling thickness and joint filling area. The joint filling area values are presented as a percentage of the total fracture surface area. These filling values present the relative volume of the joint with the assumption that the maximum filling thickness for the fracture can be considered as the total thickness of the joint. The count of calcite bearing fracture surfaces is 5034, for sulphide 3945 and for the clay 7696.

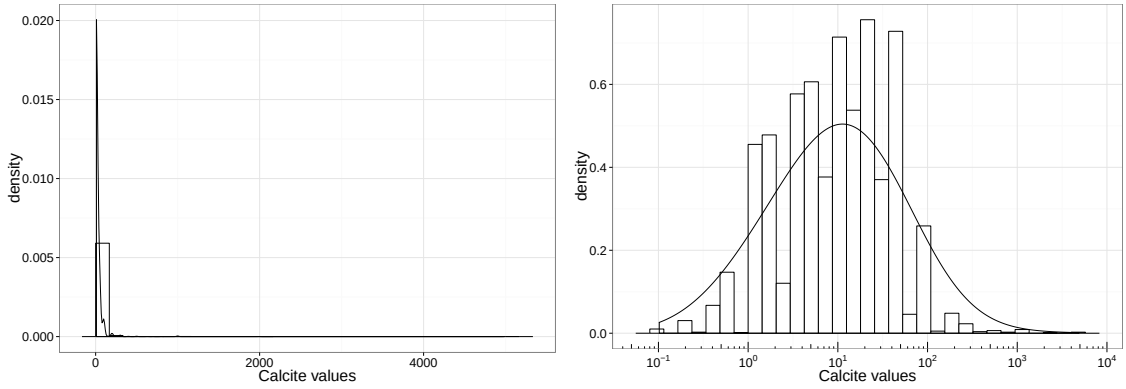

Figure 4.6: Carbonate [mm^3] histogram and kernel density plots in original (left) and logarithmic in base 10 (right) cases.

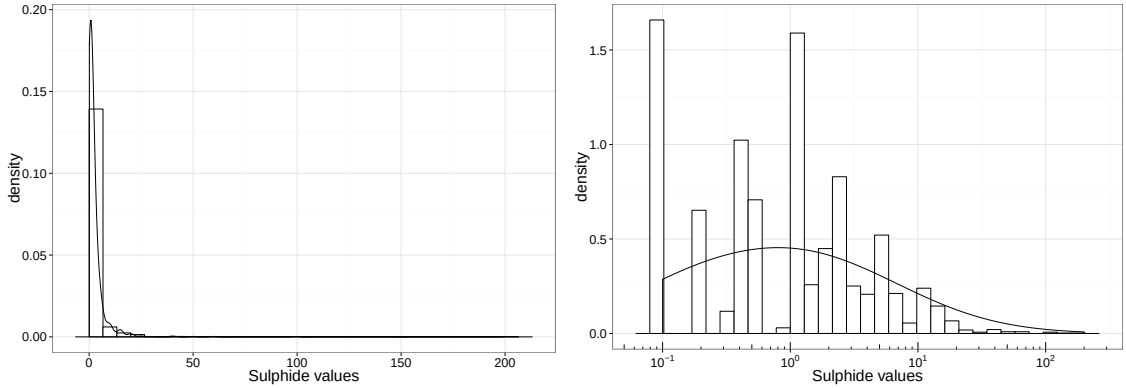

Figure 4.7: Sulphide [mm^3] histogram and kernel density plots in original (left) and logarithmic in base 10 (right) cases.

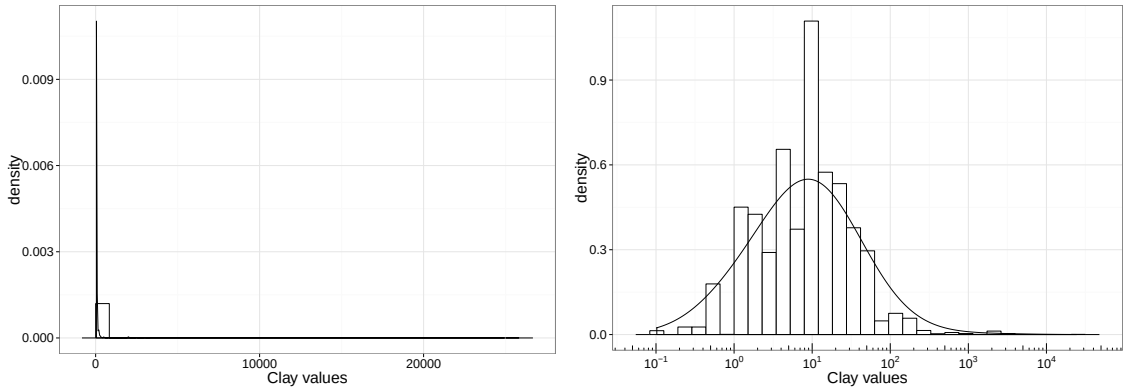

Figure 4.8: Clay [mm^3] histogram and kernel density plots in original (left) and logarithmic in base 10 (right) cases.

Following statistics have been calculated for the mineral bearing fractures. Mean value of the filling values are 25.71 mm^3, 2.43 mm^3 and 28.20 mm^3 for carbonate, sulphide and clay respectively. The corrected sample standard deviations are 116.27 mm^3 for carbonate, 6.56 mm^3 for sulphide and 313.46 mm^3 for clay. Minimum value is zero for all the groups that indicates that the joint does not have any filling that type. Maximum values are 5 000 mm^3 for carbonate, 200 mm^3 for sulphide and 25 000 mm^3 for clay. The descriptive statistics for carbonate are described in Table 4.7, for sulphide in Table 4.8 and for clay filling in Table 4.9. Based on the descriptive statistics and corresponding plots an assumption is made that the filling data are logarithmically distributed.

Table 4.7: Summary: Fracture filling Carbonate - All samples

$^{\dagger}\mu$ [mm^3]			s^2 [mm^6]	s [mm^3]	$^{\dagger\dagger}skew$	$^{\dagger\dagger}kurt$
25.71/9.25			1.35×10^4	116.27	33.19/0.30	1255.89/2.94
min [mm^3]	Q_1 [mm^3]	Me [mm^3]	Q_3 [mm^3]	max [mm^3]	IQR [mm^3]	n
0.1	4	10	25	5000	21	5034

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, IQR = interquartile range, n= sample size

† arithmetic mean / geometric mean

†† $y/y_{log_{10}}$

Table 4.8: Summary: Fracture filling Sulphide - All samples

$^{\dagger}\mu$ [mm^3]			s^2 [mm^6]	s [mm^3]	$^{\dagger\dagger}skew$	$^{\dagger\dagger}kurt$
2.43/0.81			43.00	6.56	14.24/1.39	323.57/5.18
min [mm^3]	Q_1 [mm^3]	Me [mm^3]	Q_3 [mm^3]	max [mm^3]	IQR [mm^3]	n
0.1	0.2	1	2.5	200	2.3	3945

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, $IQR =$ interquartile range, n= sample size

† arithmetic mean / geometric mean

†† $y/y_{log_{10}}$

Table 4.9: Summary: Fracture filling Clay - All samples

$^{\dagger}\mu$ [mm^3]			s^2 [mm^6]	s [mm^3]	$^{\dagger\dagger}skew$	$^{\dagger\dagger}kurt$
28.20/7.77			9.83×10^4	313.46	66.94/0.67	5244.49/4.45
min [mm^3]	Q_1 [mm^3]	Me [mm^3]	Q_3 [mm^3]	max [mm^3]	IQR [mm^3]	n
0.1	3	8	20	25000	17	7696

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, $IQR =$ interquartile range, n= sample size

† arithmetic mean / geometric mean

†† $y/y_{log_{10}}$

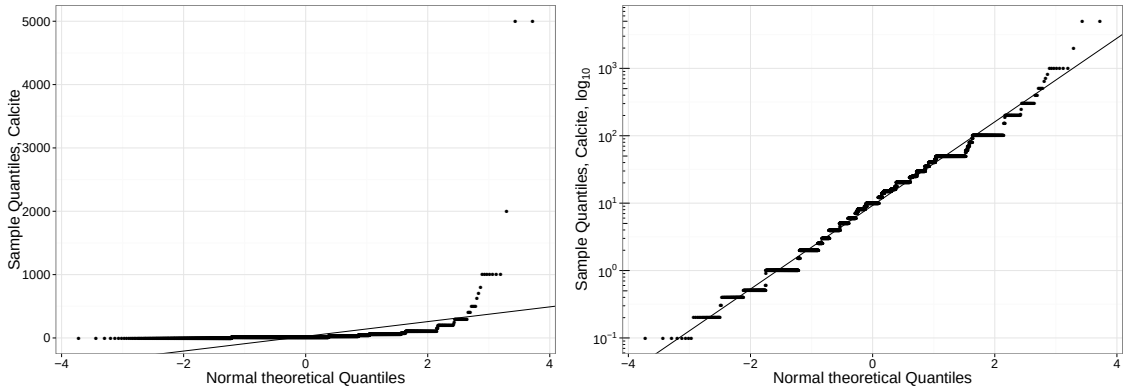

Figure 4.9: Quantile-Quantile plots of carbonate [mm^3] in original (left) and logarithmic in base 10 (right)

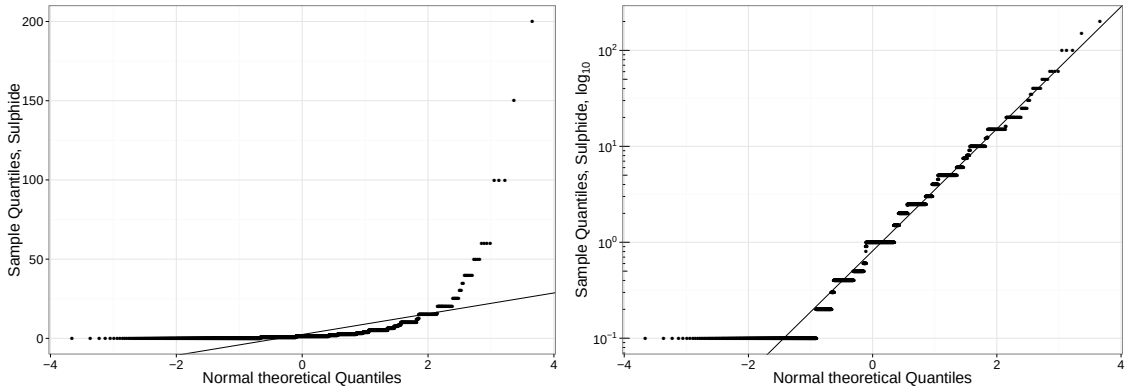

Figure 4.10: Quantile-Quantile plots of sulphide [mm^3] in original (left) and logarithmic in base 10 (right)

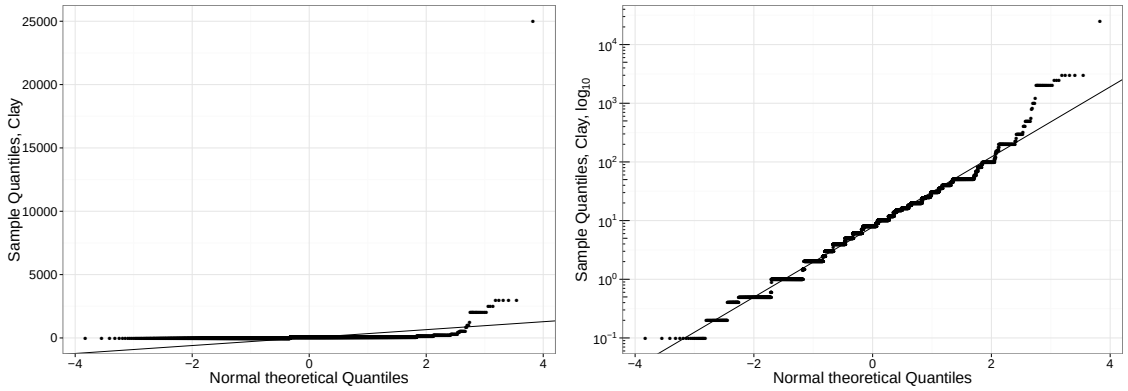

Figure 4.11: Quantile-Quantile plots clay [mm^3] in original (left) and logarithmic in base 10 (right)

4.3 Data discretisation

The data are discretized and interpolated based on the location of the transmissivity measurements. Interpolations for interval and discrete data were carried out using inverse distance weighting, IDW. The inverse distance weighting procedure is described in depth by Davis [2002, chapter 5].

4.3.1 Interval value

Interval data are sampled from the discretisation window, which is set by the modeller. The window interval is described with terms $\mathrm{L_{ref,low}}$ and $\mathrm{L_{ref,high}}$.

$$w_i = \int_{d_{i,\mathrm{low}}}^{d_{i,\mathrm{high}}} f(x)\,\mathrm{d}x \tag{45}$$

where d_{low} and d_{high} are the distances from the reference depth,L_{ref} and function $f(x)$ is the interpolation function.

$$d_{ij} = \frac{L_j - L_{\mathrm{ref}}}{L_{\mathrm{ref},i}} \quad \mathrm{for} j = \mathrm{low}, \mathrm{high} \tag{46}$$

In this thesis the interpolation function is the triangle function, that is a linear function where its sign depends on the relative location between the reference point and the mapped point.

$$f(x) = |L_{\mathrm{ref}} - x| \tag{47}$$

The different interval positions relatively to the reference point generate three different cases, where case 1 is encountered if the higher location is less than the reference point. Case 2 is encountered if the higher location value is less than sum of the window and reference point. Case 3 is encountered when the location value is bigger

than sum of the window and reference.

$$w_{i,1} = \begin{cases} \frac{1}{2}\min\left(p_{i,\text{ref}}, d_{i,\text{high}}\right)^2 - \frac{1}{2}\max\left(p_{i,\text{low}}, d_{i,\text{low}}\right)^2 & \text{for case 1} \\ -\frac{1}{2}\min\left(p_{i,\text{ref}}, d_{i,\text{high}}\right)^2 + \min\left(p_{i,\text{ref}}, d_{i,\text{high}}\right)^2 & \\ \quad + p_{i,\text{ref}}^2 - p_{i,\text{ref}} - \max\left(p_{i,\text{low}}, d_{i,\text{low}}\right)^2 & \text{for case 2} \\ -\frac{1}{2}\min\left(p_{i,\text{high}}, d_{i,\text{high}}\right)^2 + \min\left(p_{i,\text{high}}, d_{i,\text{high}}\right)^2 + d_{i,\text{low}} - \frac{1}{2}d_{i.\text{low}}^2 & \text{for case 3} \end{cases}$$

The total length of the interval is calculated as the sum of its component

$$L_w = \sum_{i=1}^{n} \left(d_{i,\text{high}} - d_{i,\text{low}}\right) \tag{48}$$

The final weighting function results from the normalising weight based on the total length.

$$w_i = \frac{w_{i,1}}{L_w} \tag{49}$$

4.3.2 Discrete value

For discrete variables the inverse distance weights are calculated based on the distance between the datum points and the reference.

$$w_{i,2} = \begin{cases} \frac{1-|L_{\text{ref},i} - L_{i,\text{point}}|}{L_{total}} & \text{If } |L_{\text{ref},i} - L_{i,\text{point}}| < 1 \\ 0 & 0 \text{ otherwise} \end{cases}$$

The weights are normalized with the total distance of the data points.

$$L_{\text{total}} = \sum_{i=1}^{n} L_{\text{ref},i} - L_{i,\text{i,point}} \tag{50}$$

4.4 Testing

Kolmogorov-Smirnov tests for transmissivity values between rock types and different joint types are performed. Based on the tests, a new grouping are made for rock- and joint types. Discretized data are compared to original dataset using IDW. Corresponding summary statistics of discritezed dataset are shown in table for each variable.

The principal factor analysis is conducted for the data with the transmissivity values. The reason for the principal factor analysis is to examine the correlation of the variables and based on the correlation to check whether the regression model indicate same kind of relationship.

Stepwise backward regression is conducted for the dataset where the logarithmic transmissivity is dependent variable. The polynomial regression is conducted based on the analyses of the data.

The regression variables are going to have a treatment based on the analyses of transmissivity values and variables. In the polynomial regression the modeller is interested only on the individual impact of parameters. Excluding grouping done previously and cross interaction between fracture surface fillings, no cross interaction are performed in the regression. Subsequently, after by the examining of the Akaike information criteria, the least influential variable is dropped. This procedure is conducted until the regression is on the satisfactory level or the polynomial regression is rejected. The polynomial regression variables are considered significant when the model AIC value less than AIC value for individual variables.

The final regression model's coefficients are analysed against the t-test and corresponding p-values are calculated. Regression model's residuals are tested against normality with Shapiro-Wilks test and corresponding Q-Q plot analysed.

5 Results

5.1 Descriptive statistics

The pooled dataset containing the data for regression and principal factor analysis are refined with discretisation procedure against the transmissivity measurements. The refined dataset excludes all the non-full groups. This reduces data to 141 points.

Mean transmissivity value is 2.36×10^{-6} $m^2\,s^{-1}$ and the corrected sample standard deviation is 1.21×10^{-5} $m^2\,s^{-1}$. Compared to the total dataset, the mean transmissivity value and the corrected standard deviation of the statistical sample are approximately same. The data are assumed to be from logarithmic distribution due to high skewness and kurtosis values, which mean that the Box-Cox transformation is applied for the values in factor analysis and regression with λ_1, $\lambda_2 = 0$. Number of observations are dropped from 176 to 141. The descriptive statistics of refined transmissivity values are described in Table 5.1.

Table 5.1: Summary statistics: Transmissivity - Refined

$^{\dagger}\mu$ [$m^2\,s^{-1}$]			s^2 [m^4/s^2]	s [$m^2\,s^{-1}$]	$^{\dagger\dagger}skew$	$^{\dagger\dagger}kurt$
$2.36\times10^{-6}/3.09\times10^{-8}$			1.47×10^{-10}	1.21×10^{-5}	8.62/0.86	8.57/3.04
min [$m^2\,s^{-1}$]	Q_1 [$m^2\,s^{-1}$]	Me [$m^2\,s^{-1}$]	Q_3 [$m^2\,s^{-1}$]	max [$m^2\,s^{-1}$]	IQR [$m^2\,s^{-1}$]	n
4.64×10^{-10}	3.11×10^{-9}	2.13×10^{-8}	1.57×10^{-7}	1.28×10^{-4}	1.54×10^{-7}	141

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, IQR = interquartile range, n= sample size

† arithmetic mean / geometric mean

†† $y/y_{log_{10}}$

The mean resistance value is 897.8 Ω m and the corrected standard deviation is over 699.3 Ω m. The single point resistance mean and corrected standard deviation for refined dataset are lower than for the original measurements. The discretisation procedure drops the number of resistance measurements from 213 074 to 141. The data has high skewness and kurtosis values, which indicates that the refined data are from logarithmic distribution. This assumption is reason for the use of Box-Cox transformation with λ_1, $\lambda_2 = 0$. The descriptive statistics for the refined resistance data are described in Table 5.2.

Table 5.2: Summary statistics: Single Point Resistance - Refined

$^{\dagger}\mu$ [Ω m]			s^2 [Ω^2 m^2]	s [Ω m]	$^{\dagger\dagger}skew$	$^{\dagger\dagger}kurt$
897.79/671.75			4.89×10^5	699.27	1.56/−0.26	5.62/2.59
min [Ω m]	Q_1 [Ω m]	Me [Ω m]	Q_3 [Ω m]	max [Ω m]	IQR [Ω m]	n
114.76	446.41	710.99	1085.65	3698.71	639.24	141

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, IQR = interquartile range, n= sample size

† arithmetic mean / geometric mean

†† $y/y_{log_{10}}$

The mean fracture frequency value for refined dataset is 5.25 m^{-1} and it is larger compared to the total dataset where the mean value was 1.87 m^{-1}. This indicates that the transmissivity measurements have been made in higher fractured rock volumes. The corrected sample deviance is slighty larger. The data has lower skewness and kurtosis values than the total dataset. Based on the high kurtosis value square root correction is performed for factor analysis and regression. The descriptive statistics for fracture frequency is described in Table 5.3.

Table 5.3: Summary statistics: Fracture frequency - Refined

$^{\dagger}\mu$ [m^{-1}]			s^2 [m^{-2}]	s [m^{-1}]	$skew$	$kurt$
5.25/††NaN			15.22	3.90	0.52	2.09
min [m^{-1}]	Q_1 [m^{-1}]	Me [m^{-1}]	Q_3 [m^{-1}]	max [m^{-1}]	IQR [m^{-1}]	n
0	1.96	4.60	8.28	13.69	6.32	141

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, $IQR =$ interquartile range, n= sample size

† arithmetic mean / geometric mean

††NaN = Not a Number

The fracture surface filling data for carbonate show that refined dataset has mean values 13.31 m^3 for all the refined points and 16.32 m^3 for fracture surfaces where carbonate occurs. The corrected standard deviation is 24.40 m^3 for all fractures and it is 26.10 m^3 for surfaces where carbonate occurs. The skewness and kurtosis values are high, which indicate that the carbonate filling comes from logarithmic distribution. The carbonate filling data includes zero values for fracture surfaces where there are no carbonates. This causes that to Box-Cox transformation is used with $\lambda_2 = 1$, which has property that the zero values map back to zero after the transformation. The number total datum points is 141 and for the points where carbonate occurs is 115. The original dataset contained 5034 fracture surfaces. The descriptive statistics are described in Table 5.4.

Table 5.4: Summary statistics: Fracture filling Carbonate - Refined

[a]Z	[†]μ [mm^3]	s^2 [mm^6]	s [mm^3]	[††]$skew$	[††]$kurt$
1	13.31/[†††]NaN	595.20	24.40	4.58/0.07	29.45/2.82
2	16.32/7.53	681.38	26.10	4.28/−0.72	25.57/4.03

[a]Z	min [mm^3]	Q_1 [mm^3]	Me [mm^3]	Q_3 [mm^3]	max [mm^3]	IQR [mm^3]	n
1	0	1.06	6.88	14.43	197.20	13.66	141
2	0.08	3.79	9.39	16.18	197.20	12.39	115

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, IQR = interquartile range, n= sample size

[a]Data 1 = Zero values included, Data 2 = Zero values excluded

[†] arithmetic mean / geometric mean

[††] $y/y_{log_{10}}$ [†††]NaN = Not a Number

The mean value of the fracture filling with sulphide is 0.96 mm^3 for all fractures and it is 1.16 mm^3 for points that contain sulphide. The median value of the refined data is 0.25 mm^3 for with zero values and 0.53 mm^3 without zeros. The original data set has higher mean value than the refined dataset. The corrected sample standard deviation is 2.06 mm^3 for all points and 2.21 mm^3 for points containing sulphide. The skewness and kurtosis values are high which indicate that the sample data are from logarithmic distribution. The Box-Cox transformation with $\lambda_2 = 1$ is applied on the data. The transformation maps zeros back to zeros. The number of occurrences is dropped from 3 945 measurements to 123 when the zero values are excluded. The descriptive statistics for fracture filling with sulphide content is described in Table 5.5.

Table 5.5: Summary statistics: Fracture filling Sulphide - Refined

[a]Z	[†]μ [mm^3]	s^2 [mm^6]	s [mm^3]	[††]$skew$	[††]$kurt$
1	0.96/[†††]NaN	4.24	2.06	4.68/1.80	27.34/6.90
2	1.16/0.40	4.89	2.21	4.31/−0.26	23.29/2.68

[a]Z	min [mm^3]	Q_1 [mm^3]	Me [mm^3]	Q_3 [mm^3]	max [mm^3]	IQR [mm^3]	n
1	0	0.05	0.25	1.09	14.20	1.04	141
2	0.01	0.14	0.53	1.23	14.20	1.09	117

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, IQR = interquartile range, n= sample size

[a]Z 1 = Zero values included, Data 2 = Zero values excluded

[†] arithmetic mean / geometric mean

[††] $y/y_{log_{10}}$ [†††]NaN = Not a Number

The mean value of the fracture filling with clay is 13.00 mm^3 for points including zero values and 14.91 mm^3 for values without zeros. The median value for clay data is 5.43 mm^3. The mean value for original data is higher than the value of refined data. The sample corrected standard deviation is 21.26 mm^3 for all points and 22.53 mm^3 for points excluding zeroes. The skewness and kurtosis are high, which indicates that the data are logarithmically distributed. The Box-Cox transformation with $\lambda_2 = 1$ is used due to zero values in the data. The descriptive statistics are described in Table 5.6.

The number of different drillholes used in the refined dataset is 19 where the highest amount of datapoints come from OL-KR19 with 21 observations. The smallest amount of observations comes from OL-KR40, OL-KR41 and OL-KR42 with 2 observations from each. The table containing information of used drillholes is described in Table 5.7.

Table 5.6: Summary statistics: Fracture filling Clay - Refined

[a]Z	[†]μ [mm^3]	s^2 [mm^6]	s [mm^3]	[††]$skew$	[††]$kurt$
1	13.00/[†††]NaN	467.28	21.26	3.24/0.24	15.45/2.44
2	14.91/6.62	507.62	22.53	3.06/−0.14	13.94/2.90

[a]Z	min [mm^3]	Q_1 [mm^3]	Me [mm^3]	Q_3 [mm^3]	max [mm^3]	IQR [mm^3]	n
1	0	1.79	5.43	12.46	133.79	10.67	141
2	0.24	3.08	6.57	14.54	133.79	11.46	123

s^2=sample variance, s=corrected sample standard deviation, $skew$=skewness, $kurt$=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, IQR = interquartile range, n= sample size

[a]Z 1 = Zero values included, Data 2 = Zero values excluded

[†] arithmetic mean / geometric mean

[††] $y/y_{log_{10}}$

Table 5.7: Summary: Used drillholes - Refined

Hole ID	Occurrence	Hole ID	Occurrence
OL-KR1	9	OL-KR28	9
OL-KR4	4	OL-KR38	13
OL-KR10	14	OL-KR39	5
OL-KR12	5	OL-KR40	2
OL-KR19	21	OL-KR41	2
OL-KR20	8	OL-KR42	2
OL-KR22	4	OL-KR43	3
OL-KR24	5	OL-KR46	2
OL-KR25	18	OL-KR48	4
OL-KR27	11		

5.2 Lithological grouping

Original lithological data have eleven different rock types. All the transmissivity values have only one rock type in their sampling window, which means that data does not need further division. While comparing the corresponding transmissivity value between the different groups data are divided to two classes. The first class consists of veined gneiss, pegmatitic granite and tonalitic-granodioritic granite. The second class consists of mainly diatexitic gneiss and mica gneiss, but includes also the other rock types. For the regression and principal factor analyses the first class is coded with the 1 and the second class with the 0. Table 5.8 describes Kolmogorov-Smirnov test for equality between rock types and Figure 5.1 show a boxplot of the rock types with the corresponding transmissivity values.

Table 5.8: Lithological types - Kolmogorov-Smirnov tests for equality based on the corresponding transmissivity values - Refined samples

$Type_1$	$Type_2$	K-S statistic	p-value†	n_1	n_2
VGN	PGR	0.169	0.626	86	22
VGN	TGG	0.279	0.596	86	7
VGN	DGN	0.290	**0.017**	86	39
VGN	MGN	0.393	**0.014**	86	18
PGR	TGG	0.364	0.370	22	7
PGR	DGN	0.268	0.214	22	39
PGR	MGN	0.465	**0.018**	22	18
TGG	DGN	0.538	**0.040**	7	39
TGG	MGN	0.611	**0.026**	7	18
DGN	MGN	0.197	0.656	39	18

VGN = veined gneiss, PGR = pegmatitic gneiss, TGG = tonalitic-granodioritic-granitic gneiss, DGN = diatexitic gneiss, MGN = mica gneiss

† p-values less than 0.05 are bolded.

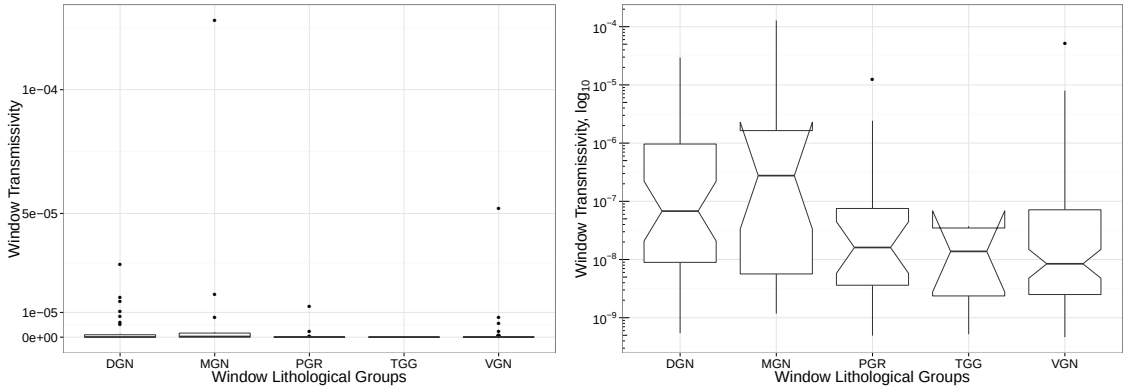

Figure 5.1: Lithological units and their corresponding transmissivity $[m^2\,s^{-1}]$ in original (left) and logaritmic in base 10 (right) cases.

VGN = veined gneiss, PGR = pegmatitic gneiss, TGG = tonalitic-granodioritic-granitic gneiss, DGN = diatexitic gneiss, MGN = mica gneiss

Table 5.9: Rock type grouping

Group number	Group members
1	VGN, PGR, TGG
2	DGN, MGN

VGN = veined gneiss, PGR = pegmatitic gneiss, TGG = tonalitic-granodioritic-granitic gneiss, DGN = diatexitic gneiss, MGN = mica gneiss

5.3 Fracture surface grouping

Fracture surfaces have eight different general groups and six geological groups, while excluding the missing values. This data are divided to 26 different combinations, which are found in sampling windows. Table 5.10 describes the occurrences of the different combinations based on the general group information. The geological group information is excluded from the analysis.

Table 5.10: Fracture type combinations within sampling windows - Refined samples

Combination	Occurrence	Combination	Occurrence
sf	32	sf + hd + sz	1
hd	1	sf + ss + fz	3
ss	2	sf + ss + pf	9
fz	1	sf + pf + fw	2
sf + hd	21	sf + hd + ss + fz	4
sf + ss	23	sf + hd + ss + pf	3
sf + fz	4	sf + hd + ss + fw	2
sf + pf	6	sf + hd +fz + pf	2
sf + fw	1	sf + hd + fz + fw	1
hd + pf	1	sf + hd + pf + fw	1
sf + hd + ss	24	sf + ss + pf + fw	2
sf + hd + fz	7	hd + ss + fz + pf	1
sf + hd + pf	6	sf + hd + ss fz + pf	1

sf = single fracture, hd = hair dyke, ss = single shear, sz = shear zone, fz = fracture zone (unweathered), fw = fracture zone (weathered), pf = paleofracture

For closer examination the general fracture type data are divided to seven groups. First four groups are extracted from Table 5.11 that are single fractures, single fractures with single shears, single fractures with hair dyke and single fractures with hair dyke and single shear. The following two groups are based on the geological interpretation. They are all the fractures containing fracture zones without weathering or fracture zones with weathered surfaces and the all the remaining groups that have paleofractures in them. The last group consist of the other remaining groups.

Fracture types are divided to four groups that are independent. The division are based on the Kolmogorov-Smirnov test for equality where transmissivity values corresponding for each sampled window are used. The first regression group consists of the windows that have only the single fractures with hair dykes, single fractures with single shear or single fractures with both hair dykes and single shears.in them. The second regression group consists of the windows that have fracture zones without weathering or fracture zones with weathered surfaces and paleofractures in them. Third group consist of groups that have single fractures and all the other groups. The highest mean values for transmissivity values are founded in the windows containing fracture zones and paleofractures. The lowest transmissivity values are in the regression group 1. The Kolmogorov-Smirnov test results are presented in Table 5.12, the fracture type groups are described in Table 5.13 and boxplot of their corresponding values are described in Figure 5.2.

Table 5.11: Fracture type grouping and corresponding transmissivity values - Refined samples

Group	†μ [$m^2 s^{-1}$]	s^2 [m^4/s^2]	s [$m^2 s^{-1}$]	*skew*	*kurt*
sf	$6.45\times10^{-7}/1.91\times10^{-8}$	9.41×10^{-12}	3.07×10^{-6}	5.34/0.86	29.66/3.58
sf+hd	$3.91\times10^{-7}/6.12\times10^{-9}/2.50\times10^{-9}$	3.02×10^{-12}	1.74×10^{-6}	4.25/2.11	19.04/7.96
sf+ss	$1.14\times10^{-7}/1.16\times10^{-8}/2.88\times10^{-9}$	7.37×10^{-14}	2.72×10^{-7}	2.99/0.62	11.41/2.33
sf+hd+ss	$1.54\times10^{-6}/1.92\times10^{-8}/3.46\times10^{-9}$	1.95×10^{-11}	4.42×10^{-6}	2.76/1.18	9.00/3.51
fz/fw	$6.37\times10^{-6}/1.23\times10^{-7}/1.05\times10^{-8}$	5.28×10^{-10}	2.30×10^{-5}	4.88/0.51	26.17/2.31
pf	$1.20\times10^{-6}/1.05\times10^{-7}/8.88\times10^{-9}$	9.62×10^{-12}	3.10×10^{-6}	3.34/0.04	13.76/2.58

Group	min [$m^2 s^{-1}$]	Q_1 [$m^2 s^{-1}$]	Me [$m^2 s^{-1}$]	Q_3 [$m^2 s^{-1}$]	max [$m^2 s^{-1}$]	IQR [$m^2 s^{-1}$]	n
sf	4.64×10^{-10}	2.97×10^{-9}	1.71×10^{-8}	4.63×10^{-8}	1.74×10^{-5}	4.33×10^{-8}	32
sf+hd	4.93×10^{-10}	1.83×10^{-9}	3.75×10^{-9}	8.35×10^{-9}	7.97×10^{-6}	6.52×10^{-9}	21
sf+ss	5.21×10^{-10}	2.44×10^{-9}	5.15×10^{-9}	3.75×10^{-8}	1.17×10^{-6}	3.50×10^{-8}	23
sf+hd+ss	7.21×10^{-10}	2.47×10^{-9}	1.05×10^{-8}	5.31×10^{-8}	1.62×10^{-5}	5.06×10^{-8}	24
fz/fw	1.41×10^{-9}	1.37×10^{-8}	8.09×10^{-8}	9.74×10^{-7}	1.28×10^{-4}	9.60×10^{-7}	32
pf	5.44×10^{-10}	1.78×10^{-8}	1.48×10^{-7}	4.32×10^{-7}	1.42×10^{-5}	4.14×10^{-7}	25

sf = only single fractures, sf+hd = only single fractures and hair dykes, sf+ss = only single fractures and single shears, sf+hd+ss = only single fracture, hair dykes and single shears, fz/fw = any windows with fault zone, pf = any left windows with paleofracture,other = all left combinations

s^2=sample variance, s=corrected sample standard deviation, *skew*=skewness, *kurt*=kurtosis, min=minimum, Q_1=first quartile, Me=Median, Q_3=third quartile max=maximum, IQR = interquartile range, n= sample size

† arithmetic mean / geometric mean

†† $y/y_{log_{10}}$

Table 5.13: Fracture type grouping

Group number	Group members
1	sf+hd, sf+ss, sf+hd+ss
2	fz/fw, pf
3	others

sf = only single fractures, sf+hd = only single fractures and hair dykes, sf+ss = only single fractures and single shears, sf+hd+ss = only single fracture, hair dykes and single shears, fz/fw = any windows with fault zone, pf = any left windows with paleofracture,other = all left combinations

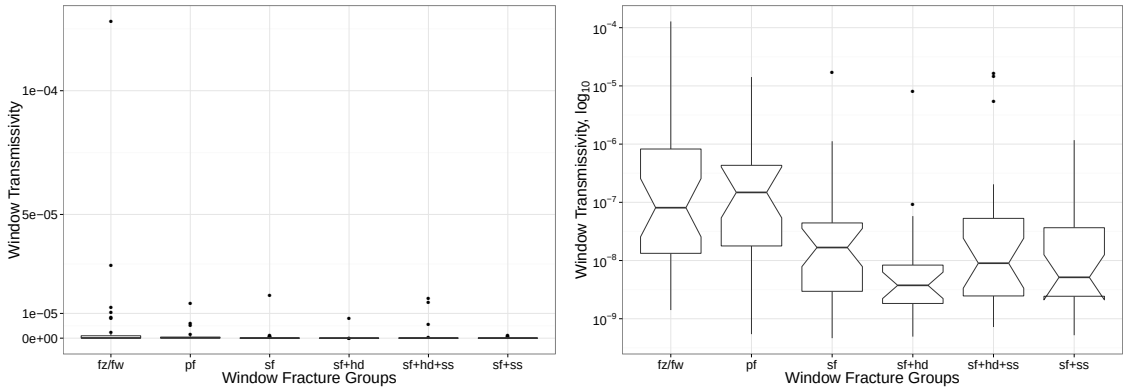

Figure 5.2: Joint surface properties, general geology, with their corresponding transmissivity [$m^2\,s^{-1}$] in original (left) and logaritmic in base 10 (right) cases.

sf = single fracture, hd = hair dyke, ss = single shear, sz = shear zone, fz = fracture zone (unweathered), fw = fracture zone (weathered), pf = paleofracture

Table 5.12: Kolmogorov-Smirnov tests - Fracture type

Type 1	Type2	D-statistic	p-value	n_1	n_2
sf	sf+hd	0.37	0.04	32	21
sf	sf+ss	0.18	0.71	32	23
sf	sf+hd+ss	0.19	0.66	32	24
sf	fz/fw	0.34	0.04	32	32
sf	pf	0.35	0.05	32	25
sf+hd	sf+ss	0.29	0.25	21	23
sf+hd	sf+hd+ss	0.36	0.09	21	24
sf+hd	fz/fw	0.61	0.00	21	32
sf+hd	pf	0.62	0.00	21	25
sf+ss	sf+hd+ss	0.13	0.97	23	24
sf+ss	fz/fw	0.37	0.04	23	32
sf+ss	pf	0.44	0.01	23	25
sf+hd+ss	fz/fw	0.35	0.05	24	32
sf+hd+ss	pf	0.42	0.02	24	25
fz/fw	pf	0.15	0.86	32	25

sf = single fracture, hd = hair dyke, ss = single shear, sz = shear zone, fz = fracture zone (unweathered), fw = fracture zone (weathered), pf = paleofracture

5.4 Principal Factor Analysis

Principal factor analysis give information about how different variables are correlating with each other. Analysed dataset does not contain any apparent number of factors, because the data are approximate distributed evenly over the nine factors. the Kaiser criterion leads to three factors that explain only 45.42 % variance. Based on the eigenvalue Table 5.14 the six highest eigenvalues contains 78.05 % of the data variance which is over 75 % that could be used as criterion for number of factors. The eight highest eigenvalues contains 94.91 % of the data variance, which is over 90 % that could be used as another criterion for number of factors.

The first eigenvalue has trace value of 19.74 % which means that the it contains 1/5 of the overall variance. The following weight of the eigenvalues decline in evenly manner and the ninth eigenvalue has a trace value 5.09 % that further indicates evenly distributed variance over different factors.

Table 5.14: Eigenvalues for correlation matrix for refined dataset

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9
Eigenvalues	2.82	1.29	1.10	0.92	0.86	0.78	0.52	0.51	0.19
$\%_{trace}$	19.74	13.33	12.35	11.30	10.93	10.40	8.45	8.41	5.09
$\sum^{\%}_{cumulative}$	19.74	33.07	45.42	56.72	67.65	78.05	86.50	94.91	100.00

The principal factor analysis in Table 5.15, contains the factor loadings for each variable. In this analysis absolute factor loading values higher than 0.50 is considered as confirming that variable is presented in a factor and all the lower values are excluded. Table 5.16 presents the processed factors using the Kaiser criterian for number of factors.

Table 5.15: Princical Factors

	F1	F2	F3	F4	F5	F6	F7	F8	F9
$\text{Trans}_{\log 10}$	-0.68	0.30	-0.06	-0.03	-0.03	-0.38	-0.54	0.03	0.02
$\text{Res}_{\log 10}$	0.48	0.51	-0.03	-0.20	0.05	0.61	-0.29	-0.10	-0.05
Frac. $\text{freq}_{\sqrt{\ }}$	-0.73	-0.26	-0.35	-0.06	-0.09	0.29	0.00	0.36	-0.22
$\text{Ca.}_{\log 10}$	-0.54	-0.13	-0.52	0.12	0.47	0.08	0.04	-0.43	-0.00
$\text{Su.}_{\log 10}$	-0.03	0.62	-0.11	0.74	0.07	0.02	0.13	0.16	-0.01
$\text{Cl.}_{\log 10}$	-0.54	-0.10	0.14	0.26	-0.68	0.21	-0.00	-0.32	0.02
Lith.	0.26	-0.65	0.30	0.47	0.23	0.20	-0.33	0.07	0.03
Ft.1	0.53	-0.17	-0.73	0.04	-0.29	0.02	-0.09	0.13	0.22
Ft.2	-0.82	0.09	0.22	-0.15	0.16	0.30	0.09	0.18	0.29

Trans = transmissivity, Res = resistivity, Frac. freq = fracture frequency, Ca = carbonate filling, Su = sulphide filling, Cl = clay filling, Lith = lithology, Ft.1 = fracture type 1, Ft.2 = fracture type 2

The variables can be divided to two factor groups with one subgroup. The first variable group includes factor 1 components with six members. The members are, in decreasing order based on the factor loadings, fracture type 2, fracture frequency, transmissivity, fracture surface clay content, fracture surface carbonate content and fracture type 1.

The first variable group has a subgroup of two factors that are fracture surface carbonate content and fracture type 1. This subgroup forms the third factor.

The second facture group includes lithology, fracture surface sulphide content and resistance values. This group forms the factor 2.

Closer examination of factor group 1 shows that the transmissivity variable has positive correlation with fracture type 2, fracture frequency, fracture surface clay content and fracture surface carbonate content. The negative correlation for transmissivity values is found with the first fracture group containing single fracture, hair dyke and single shear combinations.

The factor 3 indicates that fracture surface carbonate content has positive correlation with fracture type 1.

The factor 2 shows negative correlation with lithology and fracture surface sulphide content and resistance values. This mean that the lithology content with the second lithology group has a positive correlation with sulphide content and resistance.

Table 5.16: Princical Factors - Kaiser criteria used for number of factors

Variable	F1	F2	F3	h^2
Ft.2	-0.82			0.67
Frac. $\text{freq}_{\sqrt{\ }}$	-0.73			0.53
$\text{Trans}_{\log 10}$	-0.68			0.46
$\text{Cl.}_{\log 10}$	-0.54			0.29
$\text{Ca.}_{\log 10}$	-0.54		-0.52	0.56
Ft.1	0.53		-0.73	0.81
Lith.		-0.65		0.42
$\text{Su.}_{\log 10}$		0.62		0.38
$\text{Res}_{\log 10}$		0.51		0.26

Trans = transmissivity, Res = resistivity, Frac. freq = fracture frequency, Ca = carbonate filling, Su = sulphide filling, Cl = clay filling, Lith = lithology, Ft.1 = fracture type 1, Ft.2 = fracture type 2

5.5 Regression

The stepwise backward regression using Akaike information criteria finished with five independent variables. These variables are logarithm of the single point resistance, fracture frequency, special logarithm of the interaction variable between sulphide and clay, rock group and first fracture type set. In the final regression there exists three independent variables with negative regression coefficients that are resistance, rock group and fracture type 1. The final model consist of two positive regression coefficients that are fracture frequency and sulphide-clay interaction variable. The coefficients of the regression results are presented in Table 5.17 and the corresponding analysis of variance is presented in Table 5.18. Scatter plot of results are in Figure 5.3. The scatter plot presents that the model underestimates values from regression. The regression residuals are tested against normality with Shapiro-Wilks test and the results are presented in Table 5.19 and Q-Q plot against theoretical normal distribution is presented in Figure 5.4. The Shapiro-Wilks test indicates that the residuals are not normally distributed, but based on the Q-Q plot assumption is made that the residuals are approximately from normal distribution.

Table 5.17: Multiple linear regression model - Refined samples, final

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-6.120	0.899	-6.81	0.000
$\text{Resistance}_{\log_{10}}$	-0.537	0.273	-1.96	0.052
$\text{Fracture frequency}_{\sqrt{\ }}$	0.312	0.105	2.98	0.003
$(\text{Sulphide \& Clay interaction})_{\log_{10}}$	0.363	0.152	2.39	0.018
Lithology	-0.693	0.192	-3.60	0.000
Fracture type 1	-0.565	0.180	-3.14	0.002

Table 5.18: ANOVA table for multiple linear regression model - Refined samples, final

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
$\text{Resistance}_{\log_{10}}$	1	13.26	13.26	12.33	0.0006
$\text{Fracture frequency}_{\sqrt{\ }}$	1	16.19	16.19	15.05	0.0002
$(\text{Sulphide \& Clay interaction})_{\log_{10}}$	1	13.04	13.04	12.12	0.0007
Lithology	1	14.69	14.69	13.66	0.0003
Fracture type 1	1	10.58	10.58	9.83	0.0021
Residuals	135	145.24	1.08		

Table 5.19: Shapiro-Wilk test for residuals - Refined samples, final

W_{S-W}	p-values
0.953	9.15×10^{-5}

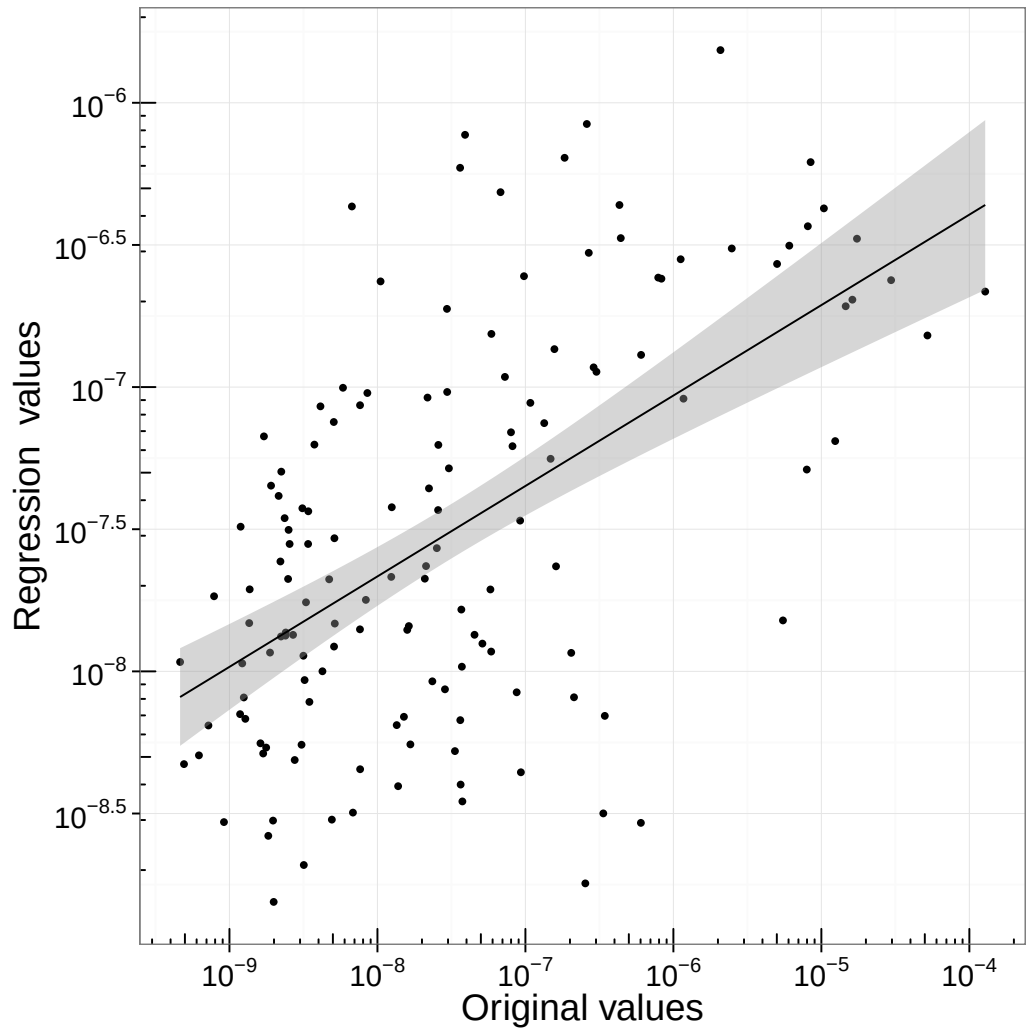

Figure 5.3: Estimated transmissivity [$m^2\,s^{-1}$] compared against original transmissivity [$m^2\,s^{-1}$]. Shaded area is 95 % confidence region.

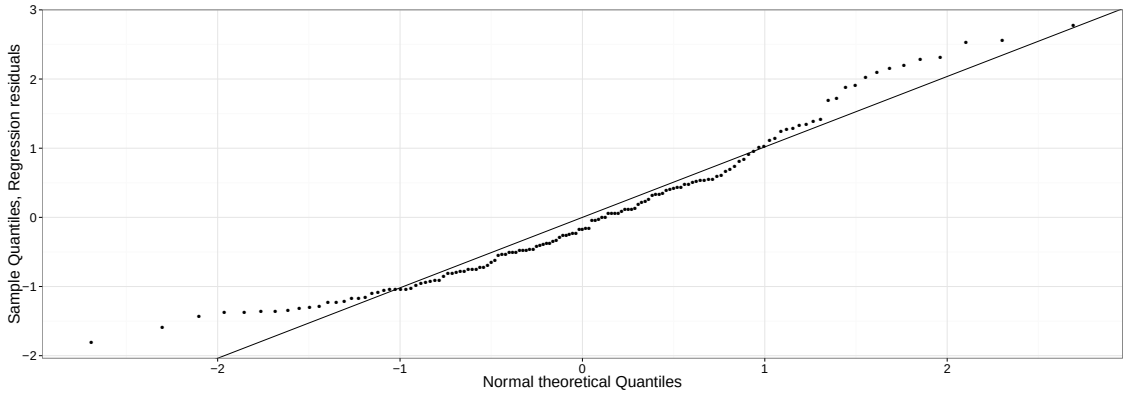

Figure 5.4: Residuals Q-Q plots against theoretical normal distribution

6 Discussion

The descriptive statistics and corresponding figures show that the transmissivity values have broad scale that can be approximated with logarithmic distribution. The fact that data after logarithmic transform are still positively skewed may be a result of resolution of measuring device. In this thesis, transmissivity data was not corrected for this error, but studies using theoretical distributions should consider using appropriate correction. The corrected distribution could be used as a parameter in modelling to give more realistic results.

The fracture frequency data being from Poisson distribution might need a better transformation due to high proportion of lower values compared to higher values as seen in Figure 4.3. The square root correction is seen acceptable for this thesis to use in principal factor analysis and linear regression.

The resistance values are clearly from logarithmic distribution. The kurtosis value is high for original and logarithmic correction, which indicates that there exists peakedness in the data set with these transformations. Even when the histogram and kernel density figure present quite symmetric plots for the resistance, the Q-Q plot in Figure 4.5 indicates that the resolution of measuring device is at its limits in the lower values. The values with the resistance higher than approximately 100 Ω m seems to follow the logarithmic distribution.

The lithological values are divided to groups based on their corresponding transmissivity value. There the results of division are satisfying for this thesis. Another way to perform the grouping could be using a priori knowledge about the geological conditions.

The fracture types are divided to groups using corresponding transmissivity measurements for individual sets that are found in the discretisation windows. This can dilute or hide the transmissivity values for highly conductive fracture types. Further dilution happens when the directional data for fractures are disregarded in discretisation procedure. These problems could be circumvented by using the directional data in the grouping and analysing more closely the fracture types and the transmissivity data.

The fracture filling mineralogy data used in this thesis consisted of three types of filling material. To further increase the accuracy of the fracture filling information other types of fillings should be included. The decision to exclude this additional data was done by modeller due to interest of specific filling materials behaviour against the transmissivity values. The used parameter for the filling materials are relative volume and this might suffer due to unknown value of real volume of the fractures.

The factor analysis was done with principal component procedure. This way the number of factors can be determined from the whole data set. Using the maximum likelihood method, factor analysis might give different and better results with correct number of factors. [Kreyszig, 2006]

The principal factor analysis indicates that there exists connection with transmissivity values and a group of variables that are fracture types, fracture frequency, fracture filling with clay and fracture filling with carbonate. The positive corre-

lation of fracture type 2 and negative correlation of fracture type 1 indicates that the fractures such as hair dyke, single shear and subgroup of single fractures are encountered more often with low values of transmissivity values. The fracture type 2 contains all the windows that have paleofractures or fracture zones. It is probable that paleofractures and fracture zones are the responsible geological structures that cause the higher conductivity values. The positive correlation of fracture frequency and transmissivity values in the factor 1 is expected due to mechanism how the groundwater moves in bedrock. The clay filling and transmissivity values have positive correlation in the factor 1 which may be result of interaction of fracture filling of clay and fracture zones that are weathered. This same reason might be behind the positive correlation of carbonate and transmissivity values. The relationship of carbonate and fracture type 1 in factor 1 is different when compared to factor 3. This cross-correlation from negative to positive may be a result of the fractures filled with carbonate material that are closed. This means that carbonate has complex relationship with transmissivity and the other variables should considered when the calcite values are used to interpret transmissivity values. The factor 1 could be interpreted to describe the fragmentation of the bedrock.

The factor 2 describes the relationship with lithology, resistance values and fracture filling of sulphide. The negative correlation of lithology between other variables tells that diatexitic and mica gneisses are positively connected to other variables. The factor 2 could be interpreted to describe the electrical resistivity or conductivity of the bedrock.

The linear regression model could be replaced by generalised linear model that could handle better the Poisson distributions and categorical data which leads to Bernoulli distributions. [Nelder and Wedderburn, 1972]

In the multiple regression results the resistance, fracture frequency, sulphide and clay interaction, lithology and fracture type 1 are survived for the last model. The resistance seems to correlate negatively with higher transmissivity values. This can be translated as the higher electrical conductance is can be detected more often with higher transmissivity values. The reason for this kind of behaviour may be consequence of hydrothermal alteration that can produce mineral with higher conductance values.

The fracture frequency correlates positively with transmissivity values, which follows the mechanism how groundwater is transferred underground. This same correlation can be seen in the factor analysis.

The sulphide clay interaction, described in page 19, has positive correlation with transmissivity values that might be associated with the hydrothermal alteration of bedrock which is more effective in fractures with high transmissivity. The interaction means that highly altered fractures are emphasized due to multiplication of sulphide and clay values.

The second fracture group is dropped from the regression, which indicates that it could be better to reduce fracture categories to two. The boxplots in Figure 5.2 show that the difference between the members of fracture type 2 and members of the single fractures that comprises the major part of the other type could be merged. Boxplot of the first fracture type set clearly indicates that the first set is having

lower transmissivity values than the combined group of others.

The negative coefficient for the rock group indicates that the first rock group is found with the lower values of transmissivity. This indicates also that diatexitic- and mica gneiss might result for higher transmissivity values.

Comparing the principal factor analysis and multiple regression shows that they give similar results but combined analyses gives better insight to variables and their behaviour. The factor 1 and regression model are very similar. The difference can be seen how the factor analysis reveals the behaviour of the carbonate filling that cannot be concluded from regression analysis alone. The regression analysis presents the relationship of lithology, resistance and a part of sulphide against the transmissivity that is not present in the principal factor analysis.

The results can have impact to the rock suitability classification [McEwen et al., 2012] of disposal holes due to correlations of lithological types, fracture types and fracture mineral assemblages with transmissivities. Figure 5.1 describing relationships between lithology and transmissivities presents that diatexitic and mica gneisses have higher transmissivities than pegmatitic granites, veined and tonalitic-granodioritic-granitic gneisses. This should be considered in rock suitability classification. Figure 5.2 presents that the surroundings of paleofractures have on average higher transmissivities than fracture type 1 group. This observation suggest that closed paleofractures with low transmissivities should be considered as potential risks. Fracture mineral assemblages has positive correlations between transmissivities in principal factor and regression analyses, which indicates that the fracture mineral assemblages have a possibility to have high transmissivities.

The combined examination of factor analysis and regression analysis has a potential to function as basis for more advanced modelling. The corrected distributions could be used to predict some of the missing datum points that would increase the number of observations used in principal factor analysis and regression analysis. Further the corrected distributions could be used to function as realistical sources for stochastical modelling [Krumbein and Dacey, 1969].

The stochastical modelling approach has a potential to include spatial awareness to modelled relationships that would lead to more advanced discretisation and correlation methods [Norberg et al., 2002]. Probabilistic methods [Guan et al., 2014] could be used to examine and simulate the risk that unwanted geological structures penetrate researched space such as fractures within the disposal holes. This can have impact on the rock suitability classification of repository that is described in McEwen et al. [2012].

Follow-up research based on the results and insights of this thesis could include linear and non-linear estimations of spatial occurrences concerning geological properties in 3D–space. The estimations could be performed with kriking [Deutsch and Journel, 1992] and co-kriking [Isaaks and Srivastava, 1990] methods or using stochastical methods such as Markov process [Krumbein and Dacey, 1969] and random field [Norberg et al., 2002] that can be used with Monte Carlo –methods [Elfeki and Dekking, 2001].

The kriking methods are linear estimation methods that estimates values between measured points [Deutsch and Journel, 1992]. The co-kriking methods uses multiple

variables to estimate values [Isaaks and Srivastava, 1990]. They are extension of kriking methods.

The Markov process or Markov chain is a stochastical modelling method that uses spatially close points to give better estimation and distribution of variables [Krumbein and Dacey, 1969]. The random field is stochastical modelling method that can have multiple continuously changing variables [Norberg et al., 2002]. The advanced simulations can give probabilities that geological properties of space are in desired range that is important aspect in rock suitability classification scheme.

7 Conclusions

The basic descriptive statistics can be helpful in exploring the properties and behaviour of the different variables. The use of different visual methods and possible hypothesis testing can result to better understanding of the dataset. The use of transformations can direct an basic analysis to correct direction and reveal the pitfalls of the measurements. In this thesis this happened for transmissivity and resistance measuring devices and the resolution of the low end values.

The descriptive statistics indicate that the transmissivity values, resistance and fracture filling data all follow logarithmic distribution. This insight lead to use of Box-Cox transformation for the previous variables. One dataset, the fracture frequency is assumed to follow the Poisson distribution due to procedure how it is gathered. The categorical variables are transformed to the Bernoulli distribution if they are grouped based on the transmissivity values.

Different fracture type values correspond to different relationships with the transmissivity. Paleofractures and fracture zones are found to correlate positively with transmissivity data.

The lithology and transmissivity have a relationship where the diatexitic and mica gneisses have positive correlation with transmissivity values.

The factor analysis gives good insight to correlations and relationships of the whole data set. By using the correct number of factors and right accepting level for factor loadings can give meaningful interpretations of the data. The factor analysis does not work in every situation, which means that interpretation should be done with caution. The analysis reveal positive correlation in factor 1 between fracture type 2, fracture frequency, transmissivity, fracture surface with clay and fracture surface with carbonate. The negative correlation is found with fracture type 1 and the previous variables. The factor analysis reveals also that fracture type 1 and carbonate content on the fracture surface have positive correlation in factor 3. This kind of behaviour can be problematic in regression analysis as well as in delineating the rock suitabilitity classification in underground storage facilities. The factor 2 describes the positive relationship with sulphide content and resistance values and lithology is found to be negatively correlated with these variables. The lithology grouping can be inverted and give a meaning that it is positively correlated with these values by changing the grouping number when the grouping is done.

The regression analysis gives more precise relationship between each of the independent variables and dependent variable. This reveals the relationship with interaction of clay and sulphide filling with the transmissivity where it is found to correlate positively as is the fracture frequency. The resistance values, lithological group 1, fracture type 1 are found to correlate negatively with the transmissivity values.

The correlations between the paleofractures, the lithological types and the fracture mineral assemblages with the transmissivities should be considered to have contributions to the rock suitability classification of disposal holes. More detailed conclusions can be made after further advanced modelling such as stochastical modelling.

In summary the statistical analysis performed properly can give general indications on the behaviour of the data set. Due to inevitable dilution from resampling a caution should be used when the result of the analysis are interpret. Different approaches can complement each other with complex relationships.

References

Ismo Aaltonen, Mari Lahti, Jon Engström, Jussi Mattila, Markku Paananen, Seppo Paulamäki, Seppo Gehör, Aulis Kärki, Turo Ahokas, Taija Torvela, and Kai Front. Geological model of the Olkiluoto site version 2.0. Posiva Working Report 70, Posiva Oy, October 2010.

Henry Ahokas, Jari Pöllänen, and Auli Kuusela-Lahtinen. Quality review of transmissivity data from Olkiluoto site – drillholes OL-KR1–KR57. Working Report Posiva 99, Posiva Oy, October 2012.

Kari Äikäs, Henry Ahokas, Annika Hagros, Eero Heikkinen, Erik Johansson, Petri Jääskeläinen, Hanna Malmlund, Paula Ruotsalainen, Pauli Saksa, Ursula Sievänen, and Pasi Tolppanen. Engineering rock mass classification of the Olkiluoto investigation site. Posiva Report 8, Posiva Oy, June 2000.

Johan Andersson, Pauliina Aalto, Ismo Aaltonen, Henry Ahokas, Susanna Aro, Malin Bomberg, Paul Degnan, Florian Eichinger, Aaron Fox, Kai Front, Andreas Gautschi, Seppo Gehör, Reija Haapanen, Matti Hakala, Lee Hartley, Jani Helin, Pirjo Hellä, John A. Hudson, Ari Ikonen, Merja Itävaara, Erik Johansson, Tuomo Karvonen, Kimmo Kemppainen, Teija Kirkkala, Lasse Koskinen, Harri Kuula, Aulis Kärki, Anne-Maj Lahdenperä, Mari Lahti, Jari Löfman, Jussi Mattila, Tim McEwen, Ferenc Mészáros, Jorge Molinero, Nicklas Nordbäck, Markku Paananen, Sami Partamies, Seppo Paulamäki, Joe Pearson, Karsten Pedersen, Ville Pietiläinen, Petteri Pitkänen, Antti Poteri, Elina Sahlstedt, Jonny Sjöberg, John A.T. Smellie, Pauli Syrjänen, Mike Thorne, Paolo Trinchero, H. Nick Waber, Tiina Vaittinen, and Mia Ylä-Mella. Olkiluoto site description 2011. Posiva Report 2, Posiva Oy, December 2012.

George E. P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

Kenneth P. Burnham and David R. Anderson. Multimodel inference, understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2):261–304, November 2004.

William J. Conover. *Practical nonparametric statistics*. Probability and statistics. Wiley, 3rd edition, 1999.

John C. Davis. *Statistics and Data Analysis in Geology*. John Wiley and Sons, Inc, 3rd edition, 2002.

Clayton V. Deutsch and André G. Journel. *GSLIB — Geostatistical Software Library and User's Guide*. Oxford University Press, 1992.

Amro Elfeki and Michel Dekking. A Markov chain model for subsurface characterization: Theory and applications. *Mathematical Geology*, 33(5):569–589, July 2001.

Jon Engström and Kimmo Kemppainen. Evaluation of the geological and geotechnical mapping procedures in use in the ONKALO access tunnel. Posiva Working Report 77, Posiva Oy, December 2008.

Aaron Fox, Kim Forchhammer, Anders Pettersson, Paul La Pointe, and Doo-Hyun Lim. Geological discrete fracture network for the Olkiluoto site, Eurajoki, Finland version 2. Posiva Report 27, Posiva Oy, June 2012.

Zhenchang Guan, Tao Deng, Yujing Jiang, Cheng Zhao, and Hongwei Huang. Probabilistic estimation of ground condition and construction cost for mountain tunnels. *Tunneling and Underground Space Technology*, 42:175–183, May 2014.

Heikki Hämäläinen. Kallioperän vedenjohtavuuden tutkimuslaitteisto. Working Report Posiva 25, Posiva Oy, June 2005.

Pirjo Hellä, Barbara Pastina, Annika Hagros, and Heini Laine. Safety case for the disposal of spent nuclear fuel at Olkiluoto — models and data for the repository system 2012. Posiva Report 1, Posiva Oy, September 2013.

Rob J. Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, November 1996.

Edward H. Isaaks and R. Mohan Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, 1990.

Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education Inc., 6th edition, 2007.

Arto Julkunen, Leena Kallio, and Pertti Hassinen. Geophysical borehole logging in pilot borehole OL-PH1 at Olkiluoto, in Eurajoki, 2004. Posiva Working Report 11, Posiva Oy, April 2004.

Erwin Kreyszig. *Advanced Engineering Mathematics*. John Wiley and Sons, Inc, 9th edition, 2006.

William C. Krumbein and Michael F. Dacey. Markov chains and embedded Markov chains in geology. *Journal of the International Association for Mathematical Geology*, 1(1):79–96, March 1969.

Johan Majapuro. Geophysical borehole logging of the boreholes KR37, KR37B and KR38, at Olkiluoto. Posiva Working Report 30, Posiva Oy, March 2006.

Jussi Mattila. A system of nomenclature for rocks in Olkiluoto. Posiva Working Report 32, Posiva Oy, June 2006.

Tim McEwen, Susanna Aro, Paula Kosunen, Jussi Mattila, Tuomas Pere, Asko Käpyaho, and Pirjo Hellä. Rock suitability classification RSC 2012. Posiva Report 24, Posiva Oy, December 2012.

Robert McGill, John. W. Tukey, and Wayne A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, February 1978.

John A. Nelder and Robert W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

Tommy Norberg, Lars Rosén, Ágnes Baran, and Sándor Baran. On modelling discrete geological structures as Markov random fields. *Mathematical Geology*, 34(1): 63–77, January 2002.

Juhani Ojala, Pasi Eilu, Pertti Turunen, Arto Julkunen, and Seppo Göher. The use of gamma spectrometry in mapping alteration zones in Olkiluoto. Posiva Working Report 64, Posiva Oy, August 2007.

Robin L. Plackett. Some theorems in least squares. *Biometrika*, 37(1):147–157, June 1950.

Jari Pöllänen. Difference flow and electrical conductivity measurements at the Olkiluoto site in Eurajoki, drillholes OL-KR44, OL-KR44B, OL-KR45, OL-KR45, OL-KR46, OL-KR47 and OL-KR48, part 1. Working Report Posiva 81, Posiva Oy, November 2009.

Eveliina Tammisto and Jorma Palmén. Database for hydraulically conductive fractures – update 2010. Posiva Working Report 12, Posiva Oy, February 2011.

Tiina Vaittinen, Henry Ahokas, Jorma Nummela, and Seppo Paulamäki. Hydrogeological structure model of the Olkiluoto site - update in 2010. Posiva Working Report 65, Posiva Oy, September 2011.